Evaluating lineup fairness: Variations across methods and measures

Jamal K. Mansour

Queen Margaret University

Jennifer L. Beaudry

Swinburne University of Technology

Natalie Kalmet

Queen's University

Michelle I. Bertrand,

University of Winnipeg

R. C. L. Lindsay

Queen's University

Author Note

Jamal K. Mansour, Memory Research Group, Centre for Applied Social Sciences, Psychology & Sociology, Queen Margaret University; Jennifer L. Beaudry, Department of Psychological Sciences, Centre for Forensic Behavioural Science, Swinburne University of Technology; Natalie Kalmet, Department of Psychology, Queen's University; Michelle I. Bertrand, Department of Criminal Justice, University of Winnipeg; R. C. L. Lindsay, Department

of Psychology, Queen's University.

This research began while Jamal K. Mansour, Jennifer L. Beaudry, and Michelle I. Bertrand were in the Department of Psychology at Queen's University. A portion of this work was presented at the 2011 meeting of the Society of Applied Research in Memory and Cognition.

This research was funded in part by a grant from the Social Sciences and Humanities Research Council (SSHRC) to R. C. L. Lindsay (Grant Number 410-09-2674) in affiliation with Queen's University. The authors would like to acknowledge the tremendous amount of work by volunteers in preparing and conducting this research.

Correspondence concerning this article should be addressed to Jamal K. Mansour, Psychology & Sociology, Queen Margaret University, Edinburgh EH21 6UU Scotland, UK. Phone: +44 (0)131 474 0 000. Fax: +44 (0)131 474 0001. Email: jmansour@qmu.ac.uk

Abstract

Triers of fact sometimes consider lineup fairness when determining the suggestiveness of an identification procedure. Likewise, researchers often consider lineup fairness when comparing results across studies. Despite their importance, lineup fairness measures have received scant empirical attention and researchers inconsistently conduct and report mock-witness tasks and lineup fairness measures. We conducted a large-scale, online experiment (N = 1010) to examine how lineup fairness measures varied with mock-witness task methodologies as well as to explore the validity and reliability of the measures. In comparison to descriptions compiled from multiple witnesses, when individual descriptions were presented in the mock-witness task, lineup fairness measures indicated a higher number of plausible lineup members but more bias towards the suspect. Target-absent lineups were consistently estimated to be fairer than target-present lineups—which is problematic because it suggests that lineups containing innocent suspects are less likely to be challenged in court than lineups containing guilty suspects. Correlations within lineup size measures and within some lineup bias measures indicated convergent validity and the correlations across the lineup size and lineup bias measures demonstrated discriminant validity. The reliability of lineup fairness measures across different descriptions was low and reliability across different sets of mock witnesses was moderate to high, depending on the measure. Researchers reporting lineup fairness measures should specify the type of description presented, the amount of detail in the description, and whether the mock witnesses viewed target-present and/or -absent lineups.

Keywords: lineup fairness, lineup bias, lineup size, eyewitness identification, mock witness

Evaluating lineup fairness: Variations across methods and measures

Lineup bias—the suspect in a lineup differing noticeably from other lineup members—negatively affects the diagnosticity of identification evidence and increases innocent suspect identifications, thereby increasing the risk of wrongful convictions (Buckhout, Figueora, & Hoff, 1974; Linds ay & Wells, 1980; Mansour, Bertrand, & Lindsay, 2013). Consequently, biased lineup procedures are discouraged (e.g., FPT Heads of Prosecution Subcommittee, 2011; Police and Criminal Evidence Act Code D, 2011; Technical Working Group, 1999). In order to detect bias, eyewitness experts often rely on the mock-witness task and specialized measures of lineup fairness.

The mock-witness task generally involves presenting a lineup and a description of a perpetrator to people who have not seen the perpetrator and then asking these people to select the lineup member who best matches the description. Lineup fairness measures derived from this task provide information about the quality and variability of lineups. If lineup fillers do not match the perpetrator's description or look very different from the suspect, the suspect is likely to draw a disproportionate number of selections. Formal measures of lineup fairness are typically categorized as reflecting lineup size or lineup bias (Malpass & Lindsay, 1999; Malpass, Tredoux, & McQuiston-Surrett, 2007). Lineup size measures indicate how many plausible lineup members are in the lineup; typically reported according to effective size (Malpass, 1981), Tredoux's *e* (Tredoux, 1998; 1999), or acceptable lineup members (Malpass & Lindsay, 1999). In contrast, lineup bias measures indicate how much a suspect stands out from other lineup members (Malpass, 1981), including the proportion of suspect selections (Doob & Kirshenbaum, 1973), functional size (Wells, Lieppe, & Ostrom, 1979), suspect bias (Doob & Kirshenbaum, 1973), defendant bias (Malpass, 1981), and binomial probability (Tredoux, 1999). Supplemental

materials 1 and 2 provide details of these measures.

Despite reasonable theoretical underpinnings for lineup fairness measures, there has been minimal exploration of best practices for assessing lineup fairness (cf. McQuisting & Malpass, 2002; Wells & Bradfield, 1999). We aim to stimulate the conversation by considering three issues. First, we assessed how variations in the mock-witness task affect lineup fairness measures—specifically, the use of target-present versus target-absent lineups and different sources of descriptions. Second, we examined the extent to which lineup fairness measures have construct validity. In the absence of a concrete benchmark, we examined the convergent and discriminant validity of the measures. Because lineup size and lineup bias are assumed to reflect different constructs, we investigated convergent validity by looking at the correlations within each construct and discriminant validity by examining correlations across constructs. Third, we explored the reliability of lineup fairness measures across different descriptions (alternate-forms reliability) and different groups of mock witnesses (inter-rater reliability).

Mock-witness Task Methodology

Variability in how mock-witness tasks are conducted may systematically influence lineup fairness measures and, thus, may have implications for the use of lineup fairness measures. We considered whether target presence or the type of descriptions used in mock-witness tasks influence lineup fairness measures.

Target Presence

Research varies in terms of whether mock-witness tasks include only target-present lineups (Humphries, Holliday, & Flowe, 2012; Mansour, Lindsay, Brewer, & Munhall, 2009), only target-absent lineups (Tredoux, Parker, & Nunez, 2007; Quinlivan et al., 2012), or both (Beresford & Blades, 2006; Clare & Lewandowsky, 2004; Parker & Ryan, 1993). Some

researchers provide no information about target presence in mock-witness tasks (e.g. Lindsay, Smith, & Pryke, 1999; McQuiston & Malpass, 2002; Meissner, Tredoux, Parker, & MacLin, 2005; Molinaro, Arndofer, & Charman, 2013). Systematic differences in lineup fairness measures due to target presence would mean that any between-study comparisons should compare only target-present to target-present and/or only target-absent to target-absent lineups. We hypothesized that target-present lineups would appear more biased than target-absent lineups because the target (as the person actually described) should better fit a witness's description than any other lineup member.

Description

The descriptions presented to mock witnesses come from a variety of sources. When evaluating real lineups, researchers typically present mock witnesses with descriptions from actual eyewitnesses (e.g., Brigham, Meissner, & Wasserman, 1999; Corey, Malpass, & McQuiston, 1999). Studies using laboratory-constructed lineups have used descriptions from one or more independent individuals in pilot testing (e.g., Lindsay, Ross, Smith, & Flanigan, 1999; Mansour et al., 2012), the eyewitness-participants themselves (e.g. Nosworthy & Lindsay, 1991), eyewitness-participants from other studies (e.g., Boyce, Lindsay, & Brimacombe, 2008), or the researchers (e.g., Gonzalez, Davis, & Ellsworth, 1995). If multiple descriptions are available, researchers may combine them in some way. A modal or consensus description uses only descriptors mentioned by some proportion of the describers (e.g., 50% by Brigham, Ready, & Spier, 1990; 25% by Beresford & Blades, 2006). An alternative would be to use any non-conflicting detail mentioned in the descriptions (i.e., a comprehensive description). It is often difficult to determine how descriptions are combined when researchers fail to report how descriptions are obtained or produced (e.g., Clare & Lewandowsky, 2004 provided no

information; Haw & Fisher, 2004 and Parker & Ryan, 1993 indicated they constructed a general description). Thus, we examined whether lineup fairness measures vary based on the type of description presented to mock witnesses. Logically, we expected that descriptions containing more details about the target would elicit greater agreement about which lineup member best matches the description (i.e., the lineup will appear less fair) than those that contain fewer details. Accordingly, we predicted that comprehensive descriptions would produce the lowest fairness ratings, followed by individual, and then consensus descriptions.

Validity of Lineup Fairness Measures

There is no easily operationalized benchmark for what should be considered a fair lineup so it is challenging to quantify the validity of lineup fairness measures. Brigham and Brandt (1992) asked students and police officers to rate lineups as fair (would use), unfair (too hard to identify suspect), or unfair (too easy to identify the suspect). They found that judgements correlated moderately to highly for effective and functional sizes, but were low for acceptable lineup members. This is the only research we are aware of that has considered validity and only three measures were evaluated. Our approach was to evaluate the convergent and discriminant validity of the variety of current measures.

Convergent Validity

Measures that reflect a similar construct will correlate highly if they have convergent validity (Garson, 2013). Strong correlations may justify tailoring lineup fairness reports according to their purpose (i.e., practice vs. research). For example, if two lineup size measures are highly correlated, researchers could choose to report the simplest-to-explain measure to lay audiences, while including the most mathematically-sound lineup size measure in journal articles. We examined the intra-dimensional correlations for lineup size and lineup bias measures

separately with the expectation that intra-dimensional correlations would be moderate to high given that the different measures are intended to measure the same construct.

Discriminant Validity

Although intuitively appealing, research has not established that lineup size and lineup bias are separate lineup fairness dimensions (Malpass & Lindsay, 1999). If the constructs are orthogonal, measures of lineup size and lineup bias should not correlate. It seems more likely, however, that the dimensions are related because they both inform the likelihood of a lineup member being selected. Malpass et al. (2007) demonstrated why: it is possible to construct a lineup with low or high effective size which has low bias but it is impossible to construct a highly biased lineup with high effective size. Assuming lineup fairness measures reflect different but related dimensions, size measures were hypothesized to correlate more highly with each other than with bias measures but to still correlate significantly. Likewise, bias measures were predicted to correlate more highly with each other than with size measures. That is, we expected that significant inter-dimensional correlations would emerge, but that they would be smaller than the intra-dimensional correlations.

Reliability of Lineup Fairness Measures

Researchers should strive to base their conclusions about the quality of lineups on reliable measures. If lineup fairness measures are to be used as evidence of the quality of real lineups, they must meet the *Daubert* (1993) standard of reliability.

Reliability Across Descriptions (Alternate-forms Reliability)

If the fairness of a lineup is a feature of the lineup itself, using different descriptions with the same lineup could be thought of as an assessment of alternate-forms reliability. Only one study we are aware of has taken this approach to date. Corey et al. (1999) manipulated whether

the description used in a mock-witness task contained distinctive descriptors (e.g., the perpetrator had squinty eyes). When the suspect did not stand out because of these descriptors, effective size and functional size were higher and defendant bias was lower than when he did stand out. Removing the distinctive information from the description reduced bias but did not increase lineup size. Based on Corey et al., we hypothesized that natural variability in descriptions from multiple witnesses would produce variability in lineup fairness measures.

Reliability Across Mock Witnesses (Interrater Reliability)

If lineup fairness measures are reliable, the specific mock witnesses involved should have little effect on the measures. To our knowledge, no one has examined the reliability of lineup fairness measures using inter-rater reliability, which is commonly used to evaluate reliability of data coding schemes (Crano & Brewer, 2002). We did not have specific hypotheses but note that a lack of strong correlations across mock-witness samples should lead researchers to question the reliability of lineup fairness measures.

Current Study

The current study examined how mock-witness methodology affects lineup fairness measures and the extent to which lineup fairness measures are valid and reliable. To do this, we collected numerous descriptions of multiple targets, created target-present and -absent lineups for each, and collected mock-witness judgments.

Method

Participants

We recruited participants ($N = 1010$) via undergraduate subject pools at two Canadian universities, through colleagues' connections, and via Mechanical Turk (https://www.mturk.com/mturk/). Most (73%) participants were female with an average age of

19.58 years (*SD* = 3.84; *Range* = 17–52). Participants' self-reported ethnicity was mixed with 56% identifying as European and 29% as Asian.

Design

We used a 3 (Description: individual, consensus, comprehensive) x 2 (Target presence: present, absent) mixed-factorial design, with description manipulated between subjects. Description indicates the type of target description participants read for the mock-witness task. Individual descriptions were from one eyewitness while consensus descriptions derived from information reported by half or more of a set of eyewitnesses, and comprehensive descriptions included all non-contradictory information reported by a set of eyewitnesses. Target presence refers to whether a lineup in the mock-witness task included the target or not.

Participants viewed description–lineup pairs for each of the 34 targets. Participants in the individual description condition were randomly assigned to receive one of 12 sets of individual descriptions (see Target Descriptions for more details). We counterbalanced target presence such that, for each description–lineup pair, half of the participants saw a target-present lineup and the rest saw a target-absent lineup.

Materials

Targets. Targets were 17 White males and 17 White females between the ages of 18 and 30. The targets varied in physical appearance, including hair color, hair length, and eye color. Distinguishing marks (e.g. moles, scars) and accessories (e.g., hairbands, earrings) were visible to participants who provided descriptions, but not to mock witnesses. Supplemental material 3 provides screen shots from the mock-crime videos of some of the targets used.

Target descriptions. An independent sample of participants (*N* = 235) were informed about the role of descriptions in eyewitness identification in the real world and laboratory

studies. We then asked them to provide the best description of targets they could. Each participant watched 7–12 targets in short, colour videos (about 30s per target, filling a 19" monitor) with each video depicting a single target facing the camera from the shoulders up.

The number of descriptions per target differed, so we randomly selected 32 descriptions per target. Research assistants coded the feature information in each description. Next, we determined how often each feature was mentioned. Finally, we generated the different types of descriptions used in the mock-witness task. For individual descriptions, we randomly selected 12 of the 32 verbatim descriptions. We created 12 sets of individual descriptions by randomly assigning each individual description for each target to a set. For consensus descriptions, we created a new description using only features (e.g., hair colour) mentioned by at least half (16) of the witnesses; the description contained the most common response (e.g., brunette) for those features. For comprehensive descriptions, we included all non-contradictory details mentioned in the description; we used the modal response for contradictory descriptions (e.g., 8 "blonde" vs. 6 "brunette" descriptors resulted in hair being described as blonde). Supplemental material 4 provides examples of descriptions used in the experiment.

The 12 individual descriptions, across targets, included a mean of 8.46 details (*SD* = 3.30, *Range* = 1–24). Descriptions included hair color (89%); height (76%); hair length and/or style (73%); eyes (63%); sex (61%); face shape and/or features (49%); ethnicity (41%); age (32%); shirt (25%); eyebrows (24%); mouth (19%); accessories (16%); facial hair (15%); glasses (15%); complexion (14%); nose (12%); teeth (12%); voice (11%); build (9%); weight (5%); posture, behavior, and/or walk (5%); sideburns (2%); shoulders, arms, and/or hands (2%); Adam's apple (1%); pants (0.5%); and/or shoes (0.5%). On average, consensus descriptions contained 4.70 details (*SD* = 1.22, *Range* = 2–7) and comprehensive descriptions contained

28.38 details (*SD* = 9.71, *Range* 14–49).

Lineups. We photographed lineup members without glasses and digitally removed distinguishing marks and accessories. Lineups were constructed using the iterative procedure (Turtle, Lindsay, & Wells, 2003). From our pool of faces (approximately 150 female and 200 male), we selected all that matched the target's ethnicity, sex, and hair color/style based on consensus among the authors. From these smaller sets, two research assistants separately chose the face they felt best matched the target. Each research assistant then chose the face they felt best matched the face selected in the previous step (i.e., the face one step removed from the target). This process was repeated (i.e., selecting a face that was the best match to the two-step-removed face, three-step-removed, etc.) until six or seven faces were chosen, depending on how many available faces matched the target's general description. Next, the authors reviewed the lineups with the research assistants to finalize the selection of lineup members. Where the two research assistants had selected different sets of lineup members, we determined which faces to use based on consensus amongst the authors. We also discarded lineup members that we agreed stood out relative to the others (e.g., a noticeably larger nose) and replaced them with the best match to the last selected face still in use.

The six selected faces constituted the target-absent lineup for a target; for lineups where seven faces were selected (n = 20), the first-selected face was discarded as per the iterative approach (Turtle et al., 2003). Target-present lineups were created by substituting the target for a randomly selected member from each of the target-absent lineups. We randomly positioned lineup members so targets did not appear in the same position more than seven times across the 34 lineups. The most selected member of the target-absent lineup was designated the innocent

suspect (worst-case scenario; Pryke, Lindsay, Dysart, & Dupuis, 2004). Supplemental material 5 contains example lineups.

Procedure

Participants viewed the consent statement and instructions on the experiment webpage. Most participated in one of the authors' laboratories and so completed the task on at least 19" monitors wherein the lineup filled most of the screen. After agreeing to participate, participants saw a description of a target next to a six-person simultaneous lineup. Participants were asked to select the person in the lineup that best matched the description. They were given as much time as they wanted to make a decision; however, a response was required to go on to the next trial. After completing 34 randomly presented trials (17 target-present, 17 target-absent; each target shown once), participants provided demographic information, indicated willingness to submit their data, and were debriefed.

Measures

We calculated three measures of lineup size (effective size, Tredoux's *e*, and acceptable lineup members) and five measures of lineup bias (proportion suspect selections, functional size, suspect bias, defendant bias, and the binomial probability of the proportion of suspect selections). Details about how to calculate each measure are given in supplemental materials 1 and 2. We used the absolute values of suspect bias and defendant bias to capture the degree of variability and to determine the magnitude of difference from chance. For acceptable lineup members, consistent with past literature (Brigham & Brandt, 1992; Lindsay et al., 1999), we specified that the criterion for a filler to be deemed acceptable was that they be selected more often than 75% of the likelihood of chance selection. For example, with a six-person lineup, the chance of any lineup member being selected is one in six (i.e., 16.67%). Given that 75% of

16.67% is 12.50%, a lineup member selected more than 12.50% of the time would be "acceptable".

The rationale for our lineup construction procedure was to ensure that lineup members approximately matched the description of the suspect. Despite our attempt to create lineups in which no lineup members stood out, not all lineup members perfectly matched each individual description. Thus, we conducted two sets of analyses: one on the full data set (n = 408 description-lineup pairs) and a second on a subset of data that included only those individual descriptions where all lineup fillers matched the description (hereafter referred to as the *complete match data subset*; n = 216 description-lineup pairs). The complete match data subset was constructed by having the first author code the number of lineup members in each target-absent lineup that matched all features mentioned in each description. Only descriptions where all six lineup members matched were used. Thus, all analyses were repeated, using only those cases where there was a match between all lineup members and the individual description the lineup was shown with. Differences in outcomes between data sets are noted in the results.

Results

Mock-witness Task Methodology

In order to determine how target presence and type of description affect lineup fairness measures, we conducted two multivariate analyses of variance (MANOVAs): One for lineup size and the other for lineup bias (see Table 1). For each target, participants in the individual description condition saw one of 12 descriptions, resulting in 12 data points per measure. We used these 12 data points to compute an average individual description score for each measure in our MANOVAs. Where our data violated analytic assumptions, we confirmed that the multivariate results did not differ across the omnibus statistics before reporting Wilk's Lambda.

We report 95% confidence intervals for Cohen's *d* (Cohen, 1992) and 90% confidence intervals for partial eta-squared (Steiger, 2004) in brackets.

Target presence. Target-present lineups appeared smaller (less fair) than target-absent lineups (see Figure 1), $F(3, 196) = 9.49$, $p < .001$, $d = 0.44$ [0.15, 0.72]. All of the univariate tests were significant: effective size, $F(1, 198) = 13.40$, $p < .001$, $d = 0.52$ [0.24, 0.80]; Tredoux's *e*, $F(1, 198) = 16.81$, $p < .001$, $d = 0.58$ [0.30, 0.86]; and acceptable lineup members, $F(1, 198) = 28.10$, $p < .001$, $d = 0.75$ [0.46, 1.04].

The results were the same for lineup bias measures (see Figure 1). Target-present lineups appeared more biased than target-absent ones, $F(5, 192) = 5.72$, $p < .001$, $d = 0.33$ [0.06, 0.61]. Significant univariate effects emerged for proportion suspect selections, $F(1, 196) = 16.05$, $p < .001$, $d = 0.56$ [0.28, 0.84]; suspect bias, $F(1, 196) = 24.49$, $p < .001$, $d = 0.70$ [0.41, 0.98]; and defendant bias, $F(1, 196) = 18.64$, $p < .001$, $d = 0.61$ [0.32, 0.89]. The effect was marginally significant for functional size, $F(1, 196) = 3.09$, $p = .08$, $d = 0.25$ [-0.03, 0.52], and not significant for binomial probability, $F(1, 196) = 0.58$, $p = .44$, $d = 0.09$ [-0.18, 0.37].

Consistent with our expectations, target-absent lineups appeared fairer than target-present lineups—for all lineup size measures and most lineup bias measures.

Description. A significant main effect emerged for the multivariate, $F(6, 394) = 18.32$, $p < .001$, $\eta_p^2 = .22$, [.14, .28], and individual univariate tests: effective size, $F(2, 198) = 62.48$, $p < .001$, $\eta_p^2 = .39$ [.28, .47]; Tredoux's *e*, $F(2, 198) = 47.10$, $p < .001$, $\eta_p^2 = .32$ [.22, .41]; and acceptable lineup members, $F(2, 198) = 14.87$, $p < .001$, $\eta_p^2 = .13$ [.05, .21]. Post hoc tests revealed that, for all lineup size measures, lineups appeared fairest when mock witnesses read individual descriptions, followed by comprehensive, and then consensus descriptions (all pairwise ps $\leq .008$; $0.41 < d$s < 1.83).

Description also had a significant multivariate effect on lineup bias measures, $F(10, 386) = 25.39$, $p < .001$, $\eta_p^2 = .40$ [.31, .45]. The univariate effect was significant for suspect bias, $F(2, 196) = 37.40$, $p < .001$, $\eta_p^2 = .28$ [.17, .37]; defendant bias, $F(2, 196) = 75.47$, $p < .001$, $\eta_p^2 = .44$ [.33, .51]; and binomial probability, $F(2, 196) = 4.47$, $p = .01$, $\eta_p^2 = .04$. Individual descriptions led to more biased ratings than consensus descriptions for suspect bias and defendant bias (ps < .001, ds ≥ 0.68 [0.34, 1.03]) and for binomial probability ($p = .006$, $d = 0.40$ [0.20, 0.60]). For binomial probability, individual descriptions also led to more biased ratings than comprehensive descriptions ($p = .02$, $d = 0.48$ [0.24, 0.72]). Consensus descriptions also produced more biased ratings than comprehensive descriptions for suspect bias ($p = .04$, $d = 0.35$ [0.17, 0.52]). The univariate effect was marginally significant for proportion suspect selections, $F(2, 196) = 2.72$, $p = 0.07$, $\eta_p^2 = .03$ [0, .08], such that individual descriptions resulted in more biased ratings than consensus descriptions ($p = .03$, $d = 0.38$ [0.19, 0.57]), but not comprehensive descriptions ($p = .65$); which did not differ ($p = .08$). The univariate effect was not significant for functional size, $F(2, 196) = 1.78$, $p = .17$, $\eta_p^2 = .02$ [0, .06].

To summarize, the results were inconsistent with our expectations that lineups would appear fairest when consensus descriptions were used and least fair when comprehensive descriptions were used. Lineups appeared largest when paired with individual descriptions and smallest when paired with consensus descriptions. Individual descriptions consistently elicited the highest bias ratings, but the effect of description was only significant for three of the five bias measures (suspect bias, defendant bias, and binomial probability).

Target presence by description interactions. The main effects of target presence and description on lineup size were qualified by a two-way interaction in the multivariate analysis, $F(6, 394) = 3.24$, $p = .004$, $\eta_p^2 = .05$ [.005, .08]. However, a significant univariate interaction

emerged only for acceptable lineup members, $F(2, 198) = 4.18$, $p = .02$, $\eta_p^2 = .04$ [.001, .10]. Post hoc analyses revealed target-absent lineups contained more acceptable lineup members than target-present lineups when paired with individual ($p < .001$, $d = 1.26$ [0.77, 1.75]) and comprehensive descriptions ($p = .008$, $d = 0.65$ [0.17, 1.13]), but not with consensus descriptions ($p = .22$, $d = 0.30$ [-0.18, 0.77]). The interaction was marginally significant for Tredoux's *e*, $F(2, 198) = 2.45$, $p = .09$, $\eta_p^2 = .02$ [0, .07], and the pattern of means was the same as for acceptable lineup members.

Similarly, the multivariate interaction was significant for lineup bias measures, $F(10, 386) = 5.90$, $p < .001$, $\eta_p^2 = .13$ [.06, .18]. The univariate interaction was significant for suspect bias, $F(2, 196) = 4.24$, $p = .02$, $\eta_p^2 = .04$ [.001, .10] and defendant bias, $F(2, 196) = 9.26$, $p < .001$, $\eta_p^2 = .09$ [.02, .16]; and marginal for binomial probability, $F(2, 196) = 2.98$, $p = .053$, $\eta_p^2 = .03$ [< .001, .08]. Follow-up analyses indicated that, when individual descriptions were used, target-present lineups were significantly more biased than target-absent lineups for suspect bias (ps < .001, $d = 1.27$ [0.75, 1.79]) and defendant bias ($p < .001$, $d = 1.44$ [0.90, 1.97]). Target-present and -absent lineups did not significantly differ on measures of suspect bias or defendant bias when mock witnesses viewed either consensus (ps > .13) or comprehensive descriptions (ps > .06). In contrast, for binomial probability, target-present lineups were more biased than target-absent lineups when paired with comprehensive descriptions ($p = .02$, $d = 0.58$ [0.09, 1.06]). All other comparisons were not significant (ps > .37).

To summarize, acceptable lineup members indicates target-absent lineups are larger than target-present lineups only if individual or comprehensive descriptions are used. For suspect bias and defendant bias, target-absent lineups appear less biased than target-present lineups only when individual descriptions are used.

Complete match data subset. For lineup size measures, the complete match data subset revealed the same main effects as the full data set. However, the omnibus target-presence by description interaction was only marginally significant, $F(6, 394) = 1.90$, $p = .08$, $\eta_p^2 = .03$ [0, .05], and was not significant for any individual measure (ps > .38).

For the bias measures, the multivariate findings were similar but the univariate results different compared to the full data set. First, the marginally significant effect of description on proportion suspect selections became significant, $F(2, 196) = 6.74$, $p = .001$, $\eta_p^2 = .06$ [.01, .14], wherein proportion suspect selections was significantly lower for individual compared to consensus descriptions ($p < .001$, $d = 0.61$ [0.58, 0.67]; other ps > .05; see Table 1). Second, the marginally significant effect of description on binomial probability became not significant (p = .24). Third, there was no longer a significant interaction for suspect bias (p = .68) or binomial probability (p = .23).

In summary, target-absent lineups appeared fairer than target-present lineups, as predicted. Contrary to our expectations, consensus descriptions did not result in the most fair ratings. Rather, consensus descriptions elicited the smallest size ratings and while they led to lower bias ratings than individual descriptions, they did not differ from comprehensive descriptions generally. Clearly lineup fairness measures vary with mock-witness task methodology.

Convergent Validity

We next considered the correlations within the lineup size and lineup bias measures, separately, to assess the convergent validity of thinking of lineup size and lineup bias as separate dimensions of lineup fairness (see Table 2). The greater the magnitude of intra-dimensional correlations, the more convergent validity we can say these measures have. Given how these

measures are calculated, lineup size measures should correlate positively with each other; proportion suspect selections, suspect bias, and defendant bias should also correlate positively (larger values indicate more bias). Functional size and binomial probability should correlate negatively with other lineup bias measures, however, as larger values indicate less bias. We considered the relationships within lineup size and lineup bias measures with Bonferroni-adjusted alphas to control for Type I error. Specifically, we used critical alphas of .017 (.05/3) for lineup size measures and .01 (.05/5) for lineup bias measures. Only individual description-lineup pairs were used in this analysis.

Full data set. Lineup size measures correlated strongly and positively for target-present ($.80 \leq r\text{s} \leq .97$, $p\text{s} < .001$) and target-absent lineups ($.75 \leq r\text{s} \leq .96$, $p\text{s} < .001$). Clearly, the three lineup size measures we examined measure a single, related construct.

Lineup bias measure correlations varied greatly ($.02 \leq |r\text{s}| \leq .79$). Proportion suspect selections, functional size, and binomial probability correlated fairly highly ($r\text{s} > |.60|$). In contrast, suspect bias correlated weakly with proportion suspect selections (.14) and moderately with defendant bias (.34) while functional size and defendant bias correlated significantly but weakly (.16). The pattern was the same for target-absent lineups but suspect bias and defendant bias correlated more strongly (.55).

Complete match data subset. The lineup size correlations were nearly identical to the full data set for target-present and -absent lineups ($.75 \leq r\text{s} \leq .98$). Most lineup bias correlations ($.01 \leq |r\text{s}| \leq .86$) were similar to the full data set; however, the correlations between suspect bias and all of the other bias measures increased ($.23 \leq |r\text{s}| \leq .86$) as compared to the full data set ($.02 \leq |r\text{s}| \leq .55$; see Table 2), regardless of target presence.

In summary, lineup size measures correlated well—indicating convergent validity. In

contrast, convergence in lineup bias measures was found only amongst proportion suspect selections, functional size, and binomial probability.

Discriminant Validity

If lineup size and lineup bias measures tap independent dimensions, correlations across measures of lineup size and lineup bias (i.e., inter-dimensional correlations) should be small or non-existent. However, lineup size logically ought to be inversely related to lineup bias (i.e., larger lineups should be less biased). The following analysis examined the extent to which lineup size and lineup bias measures appear to be orthogonally versus obliquely related. We again applied a Bonferroni correction such that $\alpha = .00625$ (.05/8).

Full data set. The directions of the correlations were in line with expectations; as the magnitude of lineup size measures increased, the magnitude of lineup bias measures decreased. With target-present lineups, the correlations generally were significant but low, with two notable exceptions. First, proportion suspect selections correlated highly with all lineup size measures ($|rs| \geq .59$). Second, defendant bias was not significantly correlated with any of the lineup size measures ($|rs| \leq .07$). The correlations of the lineup size measures with functional size, suspect bias, and binomial probability varied in magnitude from .18 to .29. For target-absent lineups, the three lineup size measures correlated moderately with proportion suspect selections ($.46 \leq |rs| \leq .36$). Binomial probability correlated significantly but weakly with effective size (.19) and Tredoux's *e* (.16, $p = .001$), while acceptable lineup members (.10) correlated non-significantly.

Complete match data subset. Some correlations decreased in the subset, as compared to the full data set. Functional size no longer correlated significantly with lineup size measures for target-present lineups ($|rs| \leq .06$). In target-absent lineups, none of the lineup size measures significantly correlated with binomial probability ($|rs| \leq .03$).

As expected, correlations between lineup size and lineup bias measures were lower than the correlations within lineup size measures or within (some) lineup bias measures. However, some correlations were significant. Lineup size measures consistently and moderately correlated with proportion suspect selections regardless of target presence.

Reliability Across Individual Descriptions (Alternate-forms Reliability)

This analysis also used only individual description-lineup pairs. We first determined, for each target for each measure, minimum and maximum values across the 12 individual descriptions. Next, we calculated descriptive statistics using these minimum values and maximum values as data points. We describe the patterns below but direct readers to Table 3 for descriptive statistics and to supplemental materials 6-11 for boxplots.

Full data set. Lineup size varied across nearly all of the possible values (i.e., 1–6) for most targets. The measures varied considerably across lineups. Importantly, mimimum and maximum values were consistently higher—suggesting greater fairness—for target-absent than target-present lineups (see supplemental materials 6).

Similar variability emerged with measures of lineup bias. An examination of proportion suspect selections indicated that some descriptions led to a very low rate of suspect selections whereas others led to a very high rate (see supplemental materials 7). We found similar ranges for target-present and target-absent lineups (see supplemental materials 8), but generally the target-absent lineups produced lower values. The results were more extreme with functional size, which has been criticized for its ability to produce unrealistic values (Malpass & Lindsay, 1999). Suspect bias and defendant bias were also variable, although values were less extreme than functional size. Binomial probability has a restricted range (0–1) but values varied considerably such that the range of minimum values overlapped the range of maximum values—primarily

because of the variability in maxima.

Complete match data subset. Across lineup size and bias measures, the mean minima were higher and the mean maxima were lower than the full data set except that the minimum binomial probability for target-absent lineups was the same (see supplemental materials 9–11).

We expected variability in measures of lineup size and bias given different descriptions from mock-witnesses but were surprised by how dramatically the values varied—and these descriptions were not manipulated for variability but were randomly selected from a set provided by typical undergraduate participants.

Reliability Across Mock witnesses (Interrater Reliability)

To address this issue we randomly split our collected responses into two groups, calculated the lineup fairness measures for each group, and then calculated the Pearson product moment correlation between the groups for each measure (Garson, 2013). The data points for the analysis were the lineup fairness measures for each target-description pair for the particular group. Only individual description-lineup pairs were used. Table 4 presents these correlations and their associated 95% confidence intervals. Stronger correlations indicate that values of lineup fairness measures are stable across mock-witness samples. We used Bonferroni corrections to our alphas—.17 for lineup size and .01 for lineup bias—to assess whether correlations between the groups were significant. We used the same Bonferroni corrections to inferentially compare correlations (reliability) of the different lineup fairness measures.

Full data set. For lineup size measures, correlations between the groups ranged from .46 to .65 (all ps < .001) for target-present lineups and from .29 to .56 (all ps < .001) for target-absent lineups indicating moderate reliability. The reliability of acceptable lineup members was lower than the reliability for effective size in target-present, $z = 3.10$, $p = .002$, and -absent

lineups, $z = 3.37$, $p = .001$. Reliability was also lower for acceptable lineup members than Tredoux's *e* for target-present, $z = 3.58$, $p < .001$, and -absent lineups, $z = 3.75$, $p < .001$. Effective size and Tredoux's *e* did not differ in reliability for target-present or -absent lineups (*p*s > .62).

For lineup bias measures and target-present lineups, proportion suspect selections was significantly more reliable than functional size, $z = 6.98$, $p < .001$; defendant bias, $z = 8.13$, $p < .001$; and binomial probability, $z = 3.52$, $p < .001$. Suspect bias was more reliable than functional size, $z = 5.25$, $p < .001$, or defendant bias, $z = 6.40$, $p < .001$, as was binomial probability ($z = 3.49$, $p < .001$ and $z = 4.61$, $p < .001$, respectively). No other comparisons were significant (*p*s > .07).

For lineup bias measures and target-absent lineups, proportion suspect selections was significantly more reliable than functional size, $z = 4.21$, $p < .001$, and defendant bias, $z = 4.79$, $p < .001$. Binomial probability was marginally more reliable than functional size, $z = 2.80$, $p = .005$, and suspect bias was marginally more reliable than defendant bias, $z = 2.73$, $p = .0063$. All other differences were not significant (*p*s > .02).

In summary, effective size and Tredeoux's *e* were the most (and equally) reliable lineup size measures and acceptable lineup members the least reliable, though all were only moderately reliable. For lineup bias, proportion suspect selections was the most reliable and highly so, followed by suspect bias or binomial probability, which were highly reliable. Defendant bias was the least reliable, but still moderately so.

Complete match data subset. In the subset, correlations were lower for target-present ($.40 \leq r\text{s} \leq .58$) and -absent lineups ($.23 \leq r\text{s} \leq .52$), as compared to the full data set. Also, whereas effective size and Tredoux's *e* were significantly more reliable than acceptable lineup

members for target-present and -absent lineups in the full data set, there were no significant differences for target-present lineups (all *p*s ≥ .16) in the subset.

The pattern of correlations for lineup bias measures was the same as in the full data set, except that the correlations were nearly always weaker and the weakest target-present correlation was for functional size, rather than defendant bias. Also, binomial probability was no longer more reliable than functional size or defendant bias.

In summary, lineup size measures correlated moderately well across two groups of mock witnesses. Lineup bias measures were moderately to highly reliable across groups of mock witnesses, with proportion suspect selections being most reliable. Our results suggest that none of the lineup size measures and few of the lineup bias measures are sufficiently reliable to eliminate concerns about the use of the measures, particularly if the results are to be applied to real-world cases.

Discussion

Our goal was to determine the degree to which current lineup fairness measures provide valid and reliable information about lineups that were constructed to be unbiased. First, we found that lineup fairness measurement outcomes can vary considerably with the methodology and measures employed. Second, lineup size but only some lineup bias measures evidenced convergent validity while discriminant validity varied considerably across the various lineup fairness measures. Third, the reliability of lineup fairness measures across different descriptions and mock-witness samples was moderate but not impressive.

Regardless of data set, our results suggest that lineup fairness measures cannot be accepted at face value as reflecting the properties of the lineups they are used to measure. Looking across the full data set and the complete match data subset, the pattern of results varied

only slightly. Specifically, compared to the full data set, the results with the complete match data set included fewer significant univariate effects in the analyses of mock-witness methodology, slightly more convergent validity for lineup bias measures, somewhat more discriminant validity, slightly higher reliability across descriptions and slightly lower reliability across groups of mock witnesses. The fact that some correlations changed from significant with the full data set to non-significant with the complete match subset may reflect the difference in sample sizes (438 vs. 216 description–lineup pairs). On the other hand, the correlations of lineup size measures to suspect bias increased from an average of .18 (target-present) and .06 (target-absent) for the full data set to .57 and .50, respectively, for the complete match subset. This pattern supports the view that match-to-description filler selection is a critical factor for determining suspect bias. Importantly for practice, our results suggest that lineup fairness measures do not meet the *Daubert* (1993) criteria that would justify presenting them as evidence, at least for lineups constructed to be fair.

Mock-witness Task Methodology

The lineup fairness measures we tested consistently indicated that target-absent lineups increased estimates of lineup size and decreased estimates of lineup bias relative to target-present lineups. As such, lineup fairness measures are more likely to lead to challenges of the suggestiveness of lineups containing guilty suspects rather than ones containing innocents. Neither expert witnesses nor the courts want to undermine identifications from target-present lineups; however, presenting lineup fairness measures may do just that. Moreover, reliance on lineup fairness measures may reduce the opportunity of defense attorneys to challenge lineup composition for innocent suspects.

We also found that individual descriptions generally resulted in different estimates of

lineup size and bias than descriptions constructed with information from multiple individuals. Taken together, our results indicate that lineups in two studies with similar functional sizes or effective sizes may not be similarly "fair." Determining if the lineups are comparable in terms of lineup size or bias will be particularly difficult if the methodological details of the mock-witness task are not reported. We contend that the influence of target presence and description type on lineup size and bias measures, as well as the fact that target presence and description type interacted for some measures, provides further evidence that the details of mock-witness methodology must be known in order to effectively evaluate lineup fairness across experiments and situations.

Validity

Although the lineup size measures correlated well with each other—and, thus, can be said to have convergent validity—only three of the five lineup bias measures (proportion, binomial probability, and functional size) correlated highly with each other. This result may be explained by the fact that all three are based solely on the proportion of mock witnesses selecting the suspect. The fact that defendant bias and suspect bias correlate moderately with each other, but weakly with the other three measures, suggests that they may reflect a different underlying construct.

Our results also somewhat support a multi-dimensional view of lineup fairness. Some lineup size measures correlated more highly with each other than with lineup bias measures, suggesting that lineup size and lineup bias measures provide unique, if not fully independent, information.

Reliability

The reliability of lineup fairness measures could be improved. First, the reliability of

lineup fairness measures across individual descriptions was poor. Lineup size and lineup bias measures derived from different individual descriptions varied widely across the range of possible values (see supplemental materials 6–11 for further illustration). Although some variability is to be expected, these data suggest that rather than being influenced solely by the similarity of the fillers to the suspect (which we held constant), the assessment of lineup fairness also depends upon the description provided to the mock witnesses. Second, testing the exact same description-lineup pairs across two groups of mock witnesses produced considerable variance in the obtained values of lineup fairness measures. That is, for the full data set, lineup size measures were moderately reliable ($.29 \leq r\text{s} \leq .65$), whereas lineup bias measures varied from moderately (e.g., defendant bias: $r = .50$) to highly reliability (e.g., proportion suspect selections: $r = .84$). Low, or even moderate, reliability is problematic in both research and applied settings. For example, one lineup with a particular description elicited effective size values of 5.00 from one group and 3.47 from the other. Malpass (1981) argued that effective sizes less than 80% of nominal size (less than 4.8 in this case) reflect biased lineups and, thus, this lineup would be unfair given one set of mock witnesses but fair given another.

Limitations

Some limitations are worth noting. First, our targets varied little in terms of ethnicity, age, and distinctive features. Distinctiveness, as well as attractiveness, of suspects and fillers affect other judgments (e.g., Valentine & Bruce, 1986; Zebrowitz & Montepare, 2008) and, thus, may affect lineup fairness measures, particularly if mentioned in the description (Corey et al., 1999). Nonetheless, it is unlikely that additional diversity in the stimuli would reduce the variability in the lineup fairness measures—rather it may instead increase it.

Second, our analysis of validity was based on the measures themselves because there is no clear benchmark for construct validity of lineup fairness measures as of yet. Thus, our analysis of validity is based on the assumption that at least some of the lineup fairness measures are valid and provides insight into the extent that the measures are valid relative to each other, rather than objectively valid.

Third, we used the iterative method (Turtle et al., 2003) to construct our lineups and so our results may not generalize to lineups constructed using different methods (e.g., match-to-description, similarity-to-suspect, random selection). Importantly, however, our conclusions did not merit adjustment when we confined our analyses to a subset of the data where the descriptions matched all lineup members.

Future Research

Many questions about lineup fairness measures must be answered if we are to use them as indices of fairness in meaningful ways. For example, we are currently testing how well lineup fairness measures can detect variations in lineup fairness (i.e., two or five fillers that match the target's description). Future research could also take a more systematic approach when examining the correspondence between descriptions and lineup members, such as measuring or varying the prototypicality of lineup member's features to determine whether this factor affects lineup fairness (e.g., Lindsay, Martin, & Webber, 1994). For example, if the description indicates the perpetrator was blond, are mock witnesses less likely to select a lineup member with dirty blond hair than one with platinum blond hair? Descriptions from individual witnesses affected lineup fairness measures in this study and critically, the most likely to be used type of description in real-world cases, the individual description, resulted in lineups appearing largest but also most biased. Future research should thus explore how individual differences in eyewitnesses (who

provide descriptions) affect lineup fairness measures. As one of our reviewers noted, the participants that provided the descriptions for our experiment provided more information than we would expect from a real eyewitness. An important avenue to explore is how the quality and quantity of information in a description relates to the apparent fairness of a lineup assessed using that description. Likewise, researchers should test whether the approach to lineup construction (i.e., match-to-description, similarity-to-suspect, iterative selection, random selection) systematically influences the precision and reliability of values of lineup fairness measure values.

A challenging issue for theory development is why the various measures of lineup bias elicited such different results in our analysis of convergent and discriminant validity. We suggest that additional dimensions may be needed to describe lineup fairness. Perhaps rather than a single lineup bias dimension researchers should consider two dimensions—potentially the extent to which features of the suspect lead them to stand out (e.g., the most prototypical, the most attractive, etc.) and the extent to which the selected fillers cause the suspect to stand out (e.g., the proportion of good to poor fillers).

Conclusions

Malpass et al. (2007) argue that "scholars using lineups in research should evaluate and document lineup size and bias as a matter of quantifying this aspect of the stimulus materials used in their work, as a guide to replication" (p. 160). We do not question the *conceptual* importance of lineup fairness measures, in research or in real-world cases. However, our analysis indicates that current lineup fairness measures may not be fit for this purpose. Obtained values for a lineup fairness measure may be valid only for the specific description, lineup, and target used to obtain it.

Lineup fairness measures should have good psychometric properties (i.e., validity and reliability). Researchers should develop best-practice approaches for mock-witness tasks based on empirical research and acknowledge the limitations of these measures. Moreover, we should have standards for what is considered a reliable result from mock-witness tasks. Critically, when evaluating seemingly fair lineups, researchers and the courts should appreciate the variability in lineup fairness measures. If researchers include lineup fairness measures in manuscripts, they should provide a thorough description (perhaps in supplemental materials) of the mock-witness task, including the description(s) used and values obtained for target-present and target-absent lineups. We encourage researchers to refine the current mock-witness task and lineup fairness measures. In the meantime, the police should follow best-practice recommendations to construct fair lineups that are unlikely to be challenged in court.

References

Beresford, J., & Blades, M. (2006). Children's identification of faces from lineups: The effects of lineup presentation and instructions on accuracy. *Journal of Applied Psychology*, *91*, 1102–1113. doi: 10.1037/0021-9010.91.5.1102

Boyce, M. A., Lindsay, D. S., & Brimacombe, C. A. E. (2008). Investigating investigators: Examining the impact of eyewitness identification evidence on student-investigators. *Law and Human Behavior, 32*, 439-453. doi: 10.1007/s10979-007-9125-5

Brigham, J. C., & Brandt, C. C. (1992). Measuring lineup fairness: Mock-witness responses versus direct evaluations of lineups. *Law and Human Behavior*, *16*, 475–489. doi: 10.1007/BF01044619

Brigham, J.C., Meissner, C. A., & Wasserman, A. W. (1999). Applied issues in the construction and expert assessment of photo lineups. *Applied Cognitive Psychology, 13*, S73–S99. doi: 10.1002/(SICI)1099-0720(199911)13:1+3.3.CO;2-W

Brigham, J. C., Ready, D. J., & Spier, S. A. (1990). Standards for evaluating the fairness of photograph lineups. *Basic and Applied Social Psychology*, *11*, 149–163. doi: 10.1207/s15324834basp1102_3

Buckhout, R., Figueroa, D., & Hoff, E. (1975). Eyewitness identification: Effects of suggestion and bias in identification from photographs. *Bulletin of the Psychonomic Society, 6*, 71–74. doi: 10.3758/BF03333151

Clare, J., & Lewandowsky, S. (2004). Verbalizing facial memory: criterion effects in verbal overshadowing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*, 739–755. doi: 10.1037/0278-7393.30.4.739

Cohen, J (1992). A power primer. *Psychological Bulletin, 112,* 155–159. doi: 10.1037/0033-

2909.112.1.155

Corey, D., Malpass, R. S., & McQuiston, D. E. (1999). Parallelism in eyewitness and mock-witness identifications. *Applied cognitive psychology*, *13*, S41–S58. doi: 1002/(SICI)1099-0720(199911)13:1+<S41::AID-ACP632>3.0.CO;2-A

Crano, W. D., & Brewer, M. B. (2002). Principles and Methods of Social Research (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Daubert v. Merrell Dow Pharmaceuticals, 509 U.S. 579 (1993).

Doob, A. N., & Kirshenbaum, H. M. (1973). Bias in police lineups—Partial remembering. *Journal of Police Science and Administration*, *1*(3), 287–293.

FPT Heads of Prosecution Committee (2011). *The path to justice: Preventing wrongful convictions.* Retrieved from Public Prosecution Service of Canada: http://www.ppsc-sppc.gc.ca/eng/pub/ptj-spj/ptj-spj-eng.pdf

Garson, G. D. (2013). *Validity & reliability.* Asheboro: Statistical Associates Publishing.

Gonzalez, R., Davis, J., & Ellsworth, P. C. (1995). Who should stand next to the suspect? Problems in the assessment of lineup fairness. *Journal of applied psychology*, *80*, 525–531. doi: 10.1037/0021-9010.80.4.525

Haw, R. M., & Fisher, R. P. (2004). Effects of administrator-witness contact on eyewitness identification accuracy. *Journal of Applied Psychology, 89*, 1106–1112. doi: 10.1037/0021-9010.89.6.1106

Humphries, J. E., Holliday, R. E., & Flowe, H. D. (2012). Faces in Motion: Age□Related Changes in Eyewitness Identification Performance in Simultaneous, Sequential, and Elimination Video Lineups. *Applied Cognitive Psychology*, *26*, 149–158. doi: 10.1002/acp.1808

Lindsay, R.C.L., Martin, R., & Webber, L. (1994). Default values in eyewitness descriptions: A problem for the match-to-description lineup foil selection strategy. *Law & Human Behavior*, *18*, 527–541. doi: 10.1007/BF01499172

Lindsay, R. C., Ross, D. F., Smith, S. M., & Flanigan, S. (1999). Does race influence measures of lineup fairness? *Applied Cognitive Psychology*, *13*, S109–S119. doi: 10.1002/(SICI)1099-0720(199911)13:1+<S109::AID-ACP690>3.0.CO;2-4

Lindsay, R. C. L., Smith, S. M., & Pryke, S. (1999). Measures of lineup fairness: Do they postdict identification accuracy? *Applied Cognitive Psychology, 13*, S93–S107. doi: 10.1002/(SICI)1099-0720(199911)13:1+3.3.CO;2-O

Lindsay, R.C.L., & Wells, G.L. (1980). What price justice? Exploring the relationship of lineup fairness to identification accuracy. *Law & Human Behavior*, *4*, 303–313. doi: 10.1007/BF01040622

Malpass, R. S. (1981). Effective size and defendant bias in eyewitness identification lineups. *Law and Human behavior*, *5*, 299–309. doi: 10.1007/BF01044945

Malpass, R. S., & Lindsay, R. C. L. (1999). Measuring line-up fairness. *Applied Cognitive Psychology, 13*, S1–S7. doi:10.1002/(SICI)1099-0720(199911)13:1+<S1::AID-ACP678>3.0.CO;2-9

Malpass, R. S., Tredoux, C. G., & McQuiston-Surrett, D. (2007). Lineup construction and lineup fairness. In R. C. L. Lindsay, D. F. Ross, J. D. Read, & M. P. Toglia (Eds.), *Handbook of Eyewitness Psychology: Memory for People* (pp. 155–178). Mahwah, NJ: Lawrence Erlbaum and Associates.

Mansour, J. K., Lindsay, R. C. L., Brewer, N., & Munhall, K. G. (2009). Characterizing visual behaviour in a lineup task. *Applied Cognitive Psychology*, *23*, 1012–1026.doi:

10.1002/acp.1570

Mansour, J. K., Beaudry, J. L., Bertrand, M. I., Kalmet, N., Melsom, E., & Lindsay, R. C. L. (2012). Impact of disguise on identification decision and confidence with simultaneous and sequential lineups. *Law and Human Behavior, 36,* 513–526. doi: 10.1037/h0093937

Mansour, J. K., Bertrand, M. I., & Lindsay, R. C. L. (2013, March). *What Might be Missed and Noticed? Novel Biases in Lineup Construction.* Paper at the American Psychology-Law Society, Portland, OR, USA.

McQuiston, D. E., & Malpass, R. S. (2002). Validity of the mock witness paradigm: Testing the assumptions. *Law and Human Behavior, 26*, 439–453. doi: 10.1023/A:1016383305868

Meissner, C. A., Tredoux, C. G., Parker, J. F., & MacLin, O. H. (2005). Investigating the phenomenological basis for eyewitness decisions in simultaneous and sequential lineups: A dual-process signal detection theory analysis. *Memory & Cognition*, *33*, 783–792. doi: 10.3758/BF03193074

Molinaro, P. F., Arndorfer, A., & Charman, S. D. (2013). Appearance-change instruction effects on eyewitness lineup identification accuracy are not moderated by amount of appearance change. *Law and Human Behavior*, *37*, 432–440. doi: 10.1037/lhb0000049

Nosworthy, G. J., & Lindsay, R. C. (1991). Does nominal lineup size matter?. *Journal of Applied Psychology*, *75*, 358–361. doi: 10.1037/0021-9010.75.3.358

Parker, J. F., & Ryan, V. (1993). An attempt to reduce guessing behavior in children's and adults' eyewitness identifications. *Law and Human Behavior*, *17*, 11–26. doi: 10.1007/BF01044534.

Police and Criminal Evidence Act 1984 Code D: Code of practice for the exercise by police statutory powers to identify persons. (March, 2011). London: Home Office Communication Directorate. Retrieved from

https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/253831/pace-code-d-2011.pdf

Pryke, S., Lindsay, R. C. L., Dysart, J. E., & DuPuis, P. (2004). Multiple independent identification decisions: A method of calibrating eyewitness identification accuracy. *Journal of Applied Psychology*, *89*, 73–84. doi: 10.1037/0021-9010.89.1.73

Quinlivan, D. S., Neuschatz, J. S., Cutler, B. L., Wells, G. L., McClung, J., & Harker, D. L. (2012). Do pre-admonition suggestions moderate the effect of unbiased lineup instructions? *Legal and Criminological Psychology*, *17*, 165–176.doi: 10.1348/135532510X533554

Steiger, J. H. (2004). Beyond the *F* test: Effect size confidence intervals and tests of close fit in the analysis of variance and contrast analysis. *Psychological Methods, 9*, 164–182. doi: 10.1037/1082-989X.9.2.164

Technical Working Group for Eyewitness Evidence. (1999). *Eyewitness evidence: A guide for law enforcement.* Washington, DC: United States Department of Justice, Office of Justice Programs.

Tredoux, C. G. (1998). Statistical inference on measures of lineup fairness. *Law and Human Behavior, 22*, 217–237. doi:10.1023/A:1025746220886

Tredoux, C. (1999). Statistical considerations when determining measures of lineup size and lineup bias. *Applied Cognitive Psychology*, *13*, S9–S26. doi: 10.1002/(SICI)1099-0720(199911)13:1+3.0.CO;2-1

Tredoux, C., Parker, J. F., & Nunez, D. (2007). Predicting eyewitness identification accuracy with mock-witness measures of lineup fairness: Quality of encoding interacts with lineup format. *South African Journal of Psychology*, *37*, 207–222.

doi: 10.1177/008124630703700201

Turtle, J. W., Lindsay, R. C. L., & Wells, G. L. (2003). Best practice recommendations for eyewitness evidence procedures: New ideas for the oldest way to solve a case. *The Canadian Journal of Police and Security Services, 1,* 5–18.

Valentine, T., & Bruce, V. (1986). The effects of distinctiveness in recognising and classifying faces. *Perception*, *15*, 525–535. doi: 10.1068/p150525

Wells, G. L., & Bradfield, A. L. (1999). Measuring the goodness of lineups: Parameter estimation, question effects, and limits to the mock witness paradigm. *Applied Cognitive Psychology, 13*, S27-S37. doi: 10.1002/(SICI)1099-0720(199911)13:1+<S27::AID-ACP635>3.0.CO;2-M

Wells, G. L., Leippe, M. R., & Ostrom, T. M. (1979). Guidelines for empirically assessing the fairness of a lineup. *Law and Human Behavior, 3*, 285–293. doi:10.1007/BF01039807

Zebrowitz, L. A., & Montepare, J. M. (2008). Social psychological face perception: Why appearance matters. *Social and Personality Psychology Compass*, *2*, 1497–1517. doi: 10.1111/j.1751-9004.2008.00109.x

Table 1 *Average lineup fairness measures across type of description for target-present and -absent lineups*

Lineup shown	Lineup Size				Lineup Bias			
	Effective size	Tredoux's *e*	ALM	Proportion	Functional size	Suspect bias	Defendant bias	Binomial probability
				Full data set				
Individual descriptions								
Overall	4.58 (0.58)	4.46 (0.93)	3.37 (1.21)	.35 (.12)	3.19 (1.14)	10.10 (6.38)	6.33 (3.74)	.03 (.16)
TP	4.31 (0.66)	4.01 (1.06)	2.74 (1.24)	.40 (.15)	2.95 (1.47)	12.69 (7.50)	7.78 (4.20)	.06 (.23)
TA	4.86 (0.29)	4.92 (0.44)	4.00 (0.78)	.30 (.05)	3.42 (0.61)	7.27 (2.80)	4.70 (2.12)	.003 (.02)
Consensus descriptions								
Overall	3.10 (0.99)	2.86 (1.07)	2.44 (0.97)	.42 (.26)	5.25 (9.45)	5.33 (1.38)	2.37 (1.32)	.15 (.30)
TP	2.98 (1.06)	2.77 (1.14)	2.29 (0.90)	.47 (.27)	3.67 (3.53)	5.83 (4.38)	2.24 (1.02)	.16 (.31)
TA	3.23 (0.92)	2.95 (1.01)	2.59 (1.02)	.38 (.24)	6.82 (12.80)	4.51 (4.01)	2.34 (1.52)	.14 (.29)
Comprehensive descriptions								
Overall	3.89 (0.76)	3.67 (1.02)	2.91 (1.02)	.37 (.19)	2.78 (2.58)	3.78 (2.86)	2.06 (1.43)	.13 (.27)
TP	3.70 (0.77)	3.39 (1.01)	2.59 (.92)	.43 (.18)	3.85 (7.04)	4.49 (2.64)	2.29 (1.36)	.06 (.21)
TA	4.08 (0.70)	3.96 (0.95)	3.24 (1.02)	.30 (.17)	4.94 (4.08)	2.78 (2.58)	1.66 (1.30)	.20 (.31)
Across description types								
Overall	3.86 (1.00)	3.66 (1.20)	2.61 (1.13)	.38 (.20)	4.86 (3.66)	6.41 (5.46)	3.60 (3.12)	.10 (.26)
TP	3.66 (1.00)	3.39 (1.18)	2.54 (1.04)	.43 (.21)	3.49 (4.61)	7.69 (6.33)	4.12 (3.69)	.09 (.26)
TA	4.06 (0.96)	3.94 (1.16)	3.28 (1.10)	.33 (.17)	5.04 (7.75)	4.86 (3.66)	2.90 (2.12)	.12 (.26)
				Complete match data subset				
Individual descriptions								
Overall	4.63 (0.50)	4.60 (0.77)	3.54 (1.07)	.30 (.11)	3.84 (1.94)	5.43 (4.00)	3.32 (2.57)	.07 (.27)
TP	4.49 (0.54)	4.36 (0.89)	3.18 (1.19)	.34 (.13)	3.52 (2.16)	6.80 (4.75)	2.49 (1.73)	.05 (.19)
TA	4.76 (0.43)	4.84 (0.55)	3.91 (0.79)	.27 (.08)	4.17 (1.67)	4.06 (2.47)	4.15 (3.00)	.10 (.26)
Consensus descriptions								
Overall	3.10 (0.99)	2.86 (1.07)	2.44 (0.97)	.42 (.26)	5.25 (9.45)	5.33 (4.37)	2.37 (1.32)	.15 (.30)
TP	2.95 (1.06)	2.73 (1.13)	2.24 (0.87)	.47 (.27)	3.67 (3.53)	6.16 (4.63)	2.34 (1.52)	.16 (.31)
TA	3.24 (0.93)	2.97 (1.02)	2.58 (1.03)	.38 (.24)	6.83 (12.80)	4.51 (4.01)	2.40 (1.10)	.14 (.29)
Comprehensive descriptions								
Overall	3.89 (0.76)	3.67 (1.02)	2.91 (1.02)	.37 (.19)	4.39 (5.74)	3.78 (2.86)	2.06 (1.43)	.12 (.27)
TP	3.70 (0.78)	3.39 (1.01)	2.59 (0 .92)	.43 (.18)	3.85 (7.04)	4.76 (2.81)	1.66 (1.30)	.06 (.21)
TA	4.08 (0.70)	3.96 (0.95)	3.24 (1.02)	.30 (.17)	4.94 (4.08)	2.79 (2.58)	2.46 (1.47)	.20 (.31)
Across description types								
Overall	3.87 (0.99)	3.71 (1.19)	2.96 (1.11)	.36 (.20)	4.49 (6.44)	4.84 (3.85)	2.58 (1.93)	.12 (.27)
TP	3.72 (1.03)	3.50 (1.20)	2.67 (1.07)	.41 (.21)	3.68 (4.68)	5.90 (4.20)	3.00 (2.18)	.09 (.24)
TA	4.04 (0.94)	3.93 (1.15)	3.25 (1.09)	.31 (.18)	5.30 (7.75)	3.78 (3.14)	2.16 (1.56)	.15 (.29)

Note. The data are presented as follows: Mean (Standard deviation). ALM = Acceptable lineup members; Proportion = Proportion of suspect selections. TP = target-present, TA = target-absent.

Table 2

Correlations across lineup fairness measures for the target-present (N = 516–524) and -absent (N = 403–408) lineups in the full data set and target-present (N =101–102) and -absent (N =101–102) lineups in the complete match data subset (individual descriptions only)

Lineup fairness measure	Full Data Set							Complete Match Data Subset						
	Lineup Size		Lineup Bias					Lineup Size		Lineup Bias				
	Tredoux's *e*	ALM	Proportion	Functional size	Suspect bias	Defendant bias	Binomial Probability	Tredoux's *e*	ALM	Proportion	Functional size	Suspect bias	Defendant bias	Binomial Probability
Correlations														
Target-present lineups														
Effective size	.97**	.80**	-.67**	.22**	-.18**	-.06	.29**	.97**	.75**	-.72**	.10	-.54**	-.09	.07
Tredoux's *e*		.84**	-.68**	.20**	-.18**	-.07	.27**		.81**	-.75**	.12	-.60**	-.26*	.08
ALM			-.59**	.18**	-.19**	-.05	.22**			-.65**	.20	-.56**	-.28*	.14
Proportion				-.63**	.14*	.05	-.78**				-.54**	.82**	.29*	-.56**
Functional size					-.06	.02	.79**					-.36**	-.02	.77**
Suspect bias						.34**	-.06						.72**	-.36**
Defendant bias							-.02							.01
Target-absent lineups														
Effective size	.96**	.75**	-.44**	.06	-.07	-.08	.19**	.98**	.81**	-.55**	-.08	-.50**	-.29*	-.004
Tredoux's *e*		.79**	-.46**	.04	-.08	-.09	.16*		.83**	-.55**	-.08	-.50**	-.37**	-.03
ALM			-.36**	-.001	-.04	-.07	.10			-.51**	-.06	-.51**	-.40**	-.02
Proportion				-.65**	.14*	-.07	-.83**				-.49**	.86**	.30*	-.60
Functional size					-.02	.16*	.78**					-.23	.21	.72**
Suspect bias						.55**	-.10						.68**	-.38**
Defendant bias							.09							.15
95% CIs														
Target-present lineups														
Effective size	.96, .97	.77, .83	-.71, -.62	.14, .30	-.26, -.1	-.14, .02	.21, .37	.95, .98	.65, .82	-.80, -.62	-.10, .28	-.66, -.39	-.28, .11	-.12, .26
Tredoux's *e*		.81, .86	-.72, -.63	.12, .28	-.26, -.1	-.15, .01	.19, .35		.73, .86	-.82, -.65	-.08, .30	-.71, -.47	-.43, -.07	-.11, .27
ALM			-.64, -.53	.10, .26	-.27, -.11	-.13, .03	.14, .30			-.75, -.53	.01, .38	-.68, -.42	-.45, -.10	-.05, .33
Proportion				-.68, -.58	.06, .22	-.03, .13	-.81, -.74				-.67, -.39	.74, .87	.10, .46	-.68, -.41
Functional size					-.14, .02	-.06, .1	.76, .82					-.52, -.18	-0.22, 0.17	0.68, 0.84
Suspect bias						.26, .41	-.14, .02						.61, .80	-.52, -.18
Defendant bias							-.10, .06							-.19, .20
Target-absent lineups														
Effective size	.95, .97	.70, .79	-.51, -.36	-.04, .16	-.17, .03	-.18, .02	.09, .28	.97, .99	.74, .87	-.67, -.40	-.27, .12	-.63, -.33	-.46, -.11	-.20, .19
Tredoux's *e*		.75, .82	-.53, -.38	-.06, .14	-.18, .02	-.19, .01	.06, .25		.76, .88	-.67, -.40	-.27, .12	-.63, -.34	-.53, -.19	-.22, .16
ALM			-.44, -.27	-.10, .10	-.14, .06	-.17, .03	0, .19			-.64, -.35	-.25, .13	-.64, -.35	-.55, -.23	-.21, .17
Proportion				-.70, -.59	.04, .23	-.17, .03	-.86, -.80				-.62, -.32	.80, .90	.12, .47	-.71, -.46
Functional size					-.12, .08	.06, .25	.74, .82					-.41, -.04	.02, .39	.61, .80
Suspect bias						.48, .61	-.19, 0						.56, .77	-.53, -.20
Defendant bias							-.01, .19							-.04, .33

Note. ALM = Acceptable lineup members; Proportion = Proportion of suspect selections; ** $p < .001$, *$p < .01$.

Table 3

Descriptive statistics representing the variability in lineup fairness measures across individual descriptions.

Parameter	Effective size	Tredoux's *e*	ALM	Proportion	Functional size	Suspect bias	DB	Binomial
				Full Data Set				
Target-present Lineups								
Minimum								
Mean (SD)	1.93 (.51)	1.63 (.41)	1.21 (.41)	.09 (.08)	1.48 (.55)	.75 (.68)	.47 (.44)	0 (.03)
Range	1.11–3.32	1.12–3.06	1.00–2.00	0–.25	1.09–4.11	0–2.44	0–1.76	0–.15
Maximum								
Mean(SD)	4.72 (.39)	4.79 (.50)	4.32 (.59)	.73 (.16)	17.77 (11.92)	12.51 (3.42)	6.16 (3.67)	.83 (.27)
Range	3.86–5.50	3.79–5.79	3.00–5.00	.24–.92	4.00–38.00	3.63–17.08	2.43–25.46	.13–1.00
Target-absent Lineups								
Minimum								
Mean (SD)	2.44 (.49)	2.10 (.47)	1.53 (.51)	.05 (.04)	1.71 (.39)	.49 (.52)	.36 (.27)	0 (.004)
Range	1.25–3.56	1.18–3.46	1.00–2.00	0–.16	1.17–3.10	0–1.78	0 –.97	0–.02
Maximum								
Mean (SD)	4.84 (.29)	4.94 (.41)	4.44 (.56)	.61 (.12)	22.23 (11.24)	9.85 (2.70)	5.62 (2.20)	.96 (.08)
Range	4.21–5.37	4.13–5.59	3.00–5.00	.32–.86	6.17–38.00	3.40–15.61	3.77–16.97	.62–1.00
				Complete match data subset				
Target-Present Lineups								
Minimum								
Mean (SD)	2.56 (.83)	2.25 (.85)	1.76 (.82)	.12 (.08)	2.00 (1.15)	.93 (.88)	.67 (.71)	.03 (.12)
Range	1.11–4.63	1.12–4.57	1.00–4.00	0–.33	1.09–7.00	0–3.87	0–2.55	0–.72
Maximum								
Mean(SD)	4.65 (.45)	4.72 (.55)	4.12 (.69)	.59 (.20)	11.51 (8.92)	9.57 (4.08)	5.30 (3.93)	.70 (.33)
Range	3.69–5.48	3.60–5.69	3.00–5.00	.14–.92	3.00–35.00	1.78–17.08	1.59 –25.46	.01–1.00
Target-absent Lineups								
Minimum								
Mean (SD)	2.98 (.65)	2.65 (.71)	2.15 (.66)	.09 (.08)	2.07 (.53)	1.01 (.79)	.73 (.65)	0 (.01)
Range	1.88–4.50	1.68–4.60	1.00-3.00	0–.29	1.32–3.60	0–2.69	0–3.33	0–.08
Maximum								
Mean (SD)	4.59 (.50)	4.65 (.61)	4.03 (.87)	.51 (.12)	15.76(10.93)	7.65 (2.60)	4.38 (1.34)	.84 (.28)
Range	3.14–5.37	3.05–5.59	2.00–5.00	.28–.76	3.50–38.00	2.58–12.88	1.16–6.85	.07–1.00

Note. ALM = Acceptable lineup members; Proportion = Proportion of suspect selections; DB = Defendant bias; Binomial = binomial probability

Table 4

Correlations for lineup fairness measures (individual descriptions only) across two mock-witness samples

Lineup fairness measure	Full Data Set				Complete Match Data Subset			
	Target-present		Target-absent		Target-present		Target-absent	
	r [95% CI]	*n*	*r* [95% CI]	*n*	*r* [95% CI]	*n*	*r* [95% CI]	*n*
Effective size	.63 [.62, .65]	406	.49 [.47, .51]	406	.58 [.56, .60]	313	.45 [.42, .47]	313
Tredoux's *e*	.65 [.63, .67]	406	.51 [.49, .53]	406	.58 [.56, .60]	313	.45 [.43, .48]	313
ALM	.48 [.47, .49]	406	.29 [.28, .31]	406	.50 [.48, .50]	313	.23 [.22, .24]	313
Proportion	.84 [.83, .85]	406	.71 [.69, .73]	406	.81 [.80, .82]	313	.68 [.65, .70]	313
Functional size	.62 [.60, .64]	391	.52 [.50, .54]	340	.58 [.55, .60]	302	.54 [.51, .56]	285
Suspect bias	.80 [.79, .82]	406	.63 [.60, .65]	406	.76 [.75, .78]	313	.57 [.55, .60]	313
Defendant bias	.57 [.55, .59]	404	.50 [.48, .52]	405	.64 [.63, .66]	311	.35 [.33, .37]	313
Binomial probability	.75 [.73, .77]	406	.65 [.63, .67]	406	.68 [.66, .70]	313	.62 [.59, .64]	313

Note. Variations in sample sizes reflect the fact that the suspect was not always selected. ALM = Acceptable lineup members; Proportion = Proportion of suspect selections.

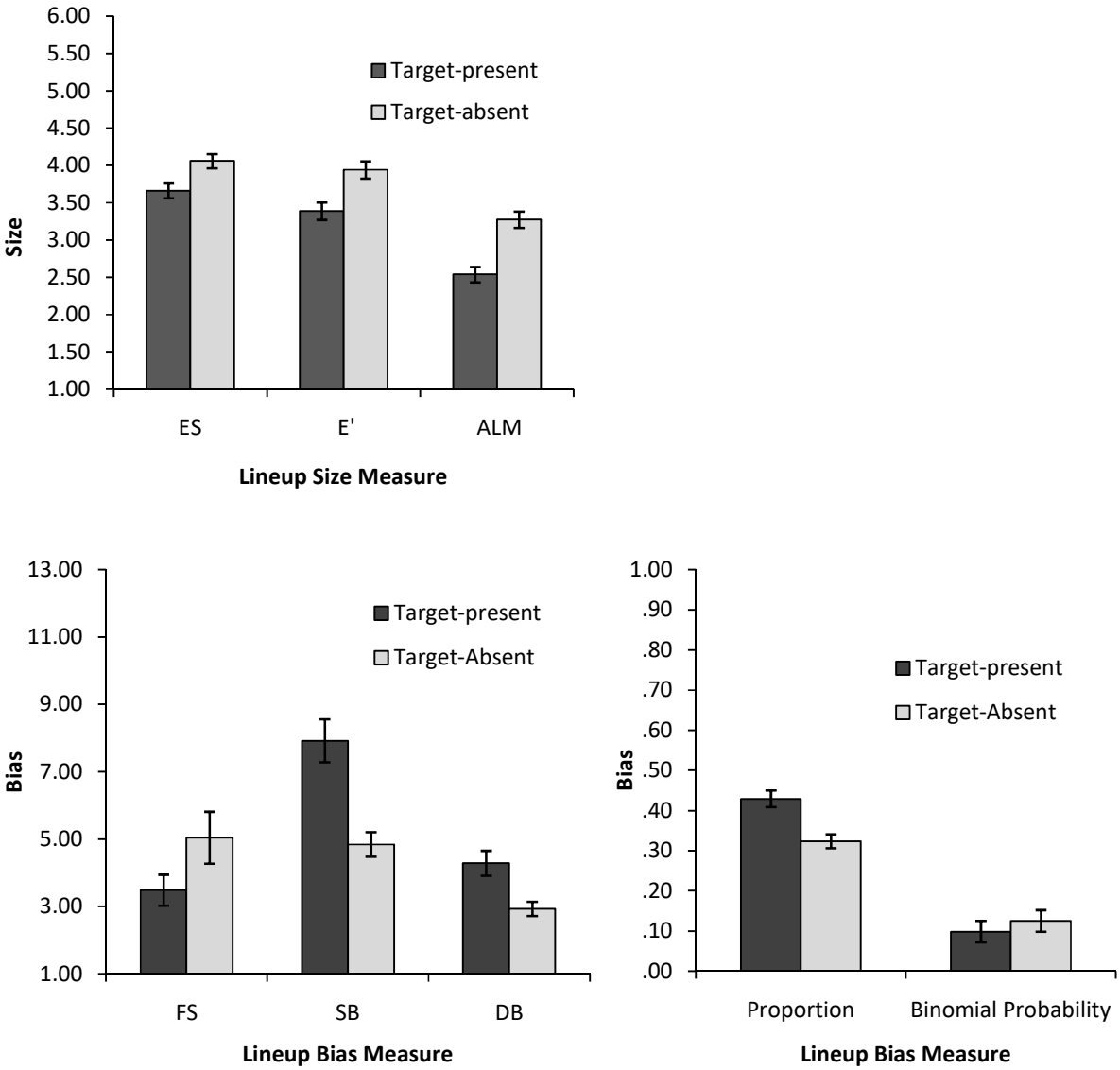


Figure 1: Average lineup size (top panel) and lineup bias (bottom two panels) for target-present and -absent lineups collapsed across target description and target. Error bars reflect standard error of the mean. ES = Effective size; E' = Tredoux's *e*; ALM = Acceptable lineup members; Proportion = Proportion of suspect selections; FS = Functional size; SB = Suspect bias; DB = Defendant bias.

Supplemental Material 1

Lineup size measures.

Measure	Description	Formula	Minimum	Maximum	Relevant citations
Effective size	How much the proportion of suspect selections differs from chance, calculated from the nominal size adjusted to include only lineup members selected at least once.	$k_a - \sum_{i=1}^{k_a} \left\lvert \frac{n_i - e_a}{2e_a} \right\rvert = k_a - \sum_{i=1}^{k_a} \left\lvert \frac{n_i - n\frac{1}{k_a}}{2n\frac{1}{k_a}} \right\rvert$	1	Nominal size	Malpass (1981)
Tredoux's *e*	This measure is similar to effective size but reflects the fact that non-selected lineup members could have been selected given more mock witnesses.	$\frac{1}{1-I} = \frac{1}{1-\left(1-\sum_{i=1}^{N}\left(\frac{n_i}{n}\right)^2\right)}$	1	Nominal size	Tredoux (1998; 1999)
Acceptable lineup members	This measure indicates how many lineup members were selected at a specified rate of chance, based on nominal size.	$\sum_{i=1}^{N} p_i > \frac{.75}{N}$	1	Nominal size	Malpass & Lindsay (1999)

Note: e_a = adjusted value for chance (one divided by the number of lineup members selected more than once and multiplied by the number of mock witnesses), k_a = number of lineup members who were selected by at least one mock witness, n = number of mock witnesses, n_i = number of mock witnesses that chose a lineup member i, N = nominal size, p_i = proportion selections of lineup member i.

Supplemental Material 2 *Lineup bias measures.*

Measure	Description	Formula	Minimum	Maximum	Relevant citations
Proportion suspect selections	The number of suspect selections divided by the number of mock witnesses.	$\frac{n_{ss}}{n}$	0	1	Brigham & Brandt (1992); Doob & Kirshenbaum (1973)
Functional size	The inverse of proportion suspect selections.	$\frac{1}{p_{ss}}$	1	Infinity	Wells, Leippe, & Ostrom (1979)
Suspect bias	A z-score comparing proportion of suspect selections to chance; i.e., the inverse of nominal size.	$\frac{p_{ss} - \frac{1}{N}}{\sigma} = \frac{p_{ss} - \frac{1}{N}}{\sqrt{\frac{\frac{1}{N}\left(1 - \frac{1}{N}\right)}{n}}}$	Negative infinity	Positive infinity	Doob & Kirshenbaum (1973)
Defendant bias	A z-score comparing the proportion of suspect selections to chance using the inverse of effective size as the estimate of chance.	$\frac{p_{ss} - \frac{1}{ES}}{\sigma} = \frac{p_{ss} - \frac{1}{ES}}{\sqrt{\frac{\frac{1}{ES}\left(1 - \frac{1}{ES}\right)}{n}}}$	Negative infinity	Positive infinity	Malpass (1981); Malpass & Lindsay (1999)
Binomial probability	The probability of the proportion of suspect selections occurring using binomial distribution (rather than the normal distribution).	$\binom{n}{n_{ss}}\frac{1^{n_{ss}}}{N} \cdot \frac{(N-1)^{n-n_{ss}}}{N}$	0	1	Tredoux (1999)

Note: n = number of mock witnesses, n_{ss} = number of suspect selections, N = nominal size, p_{ss} = proportion of suspect selections, σ is the standard deviation of the sampling distribution for the proportion of suspect selections, ES = effective size

Supplemental Material 3

Screenshots of videos presented to participants asked to give descriptions of targets.

Note: The individuals whose faces appear here gave consent for the use of their likenesses.

Supplemental Material 4

Example descriptions presented in the mock-witness task. The first description under each heading is for male target depicted in supplemental material 1 and the second description under each heading is for the female target in supplemental material 1.

Example individual descriptions (used in the complete match data subset)

“Caucasian man in his twenties, dark brown hair and eyes, hair was short but slightly longer on top - stuck up straight, long neck and rectangular shaped face”

“The person was a white female with brown hair. She wore earrings and a pink shirt.”

Example individual descriptions (NOT used in the complete match data subset)

“long face, pointed ears, brown eyes and hair, unshaven, small chin”

“Round face, woman, thin lips”

Example consensus descriptions

“Male, brown hair, short hair, long face”

“Female, brown hair”

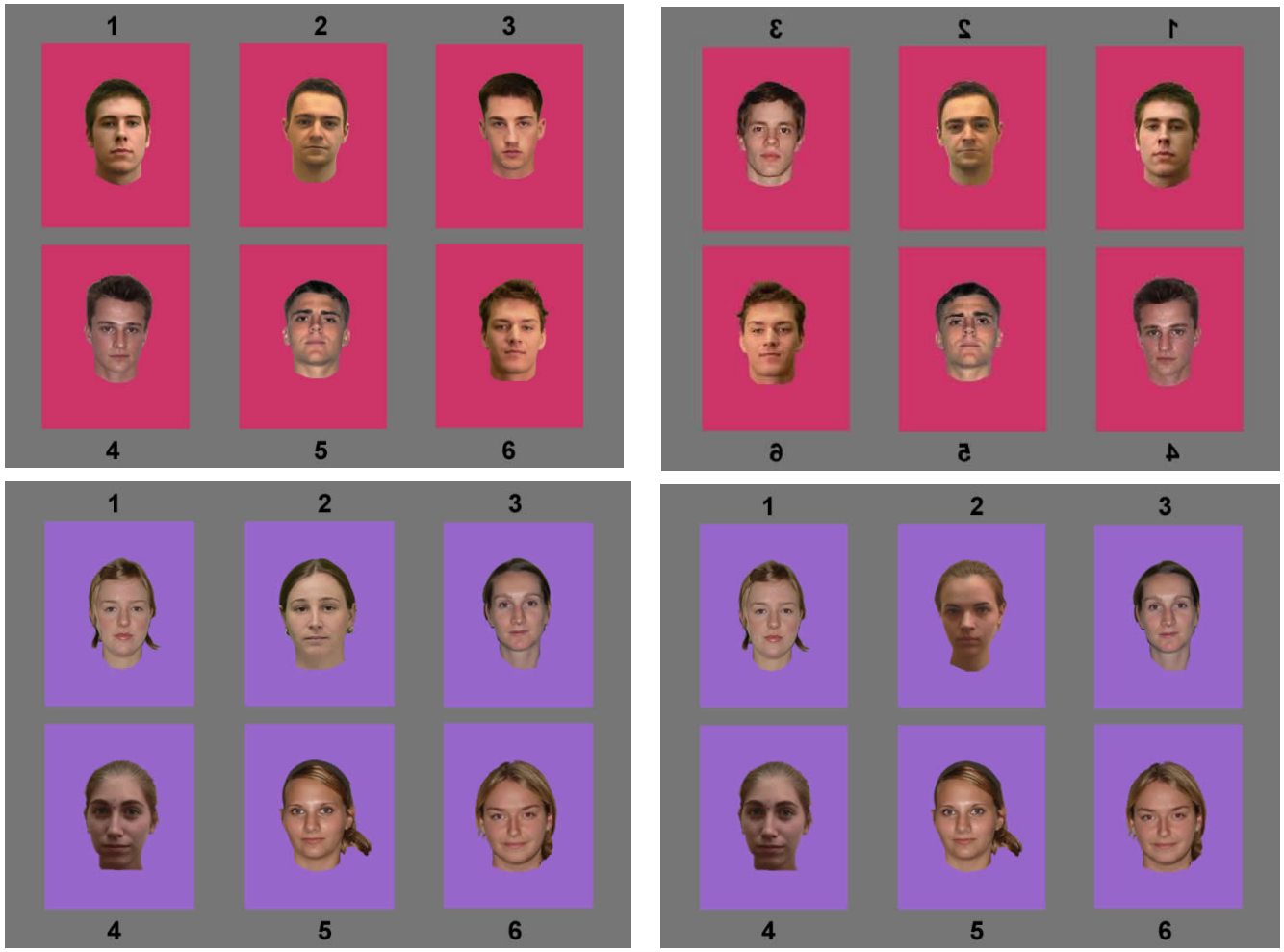
Example comprehensive descriptions

“Caucasian, Male, early to mid 20s,skinny, 140/160lbs,thin build ,brown hair, short hair, straight, possibly buzz crew cut, gelled, slight facial hair, brown, dark, shifty, possibly almond shaped or heavy lidded eyes, large nose, pouty, full bottom lip, green shirt, dark, thick eyebrows, long face, pointed ears that stick out, big forehead, small chin, medium skin tone, relaxed, tall, crooked teeth, big Adam's apple”

“Caucasian, Female, early to mid/20s,medium build ,brown hair, hair pulled back with middle part, straight, long, thin, small, brown, almond/shaped eyes; squints eyes when she speaks, blinks a lot, small nose, bigger at bottom than top, small, thin lips, pink shirt, straight, brown, thin, broad eyebrows, round face, small chin, large jaw structure, very fine features, prominent forehead, freckles, fair/skinned with freckles, too sympathetic, scared, nervous, not used to doing what they did, very analytical, straight teeth, big, chunky, square, not dangly gold earrings. Multiple piercings; one is in upper cartilage of left ear.”

Supplemental Material 5

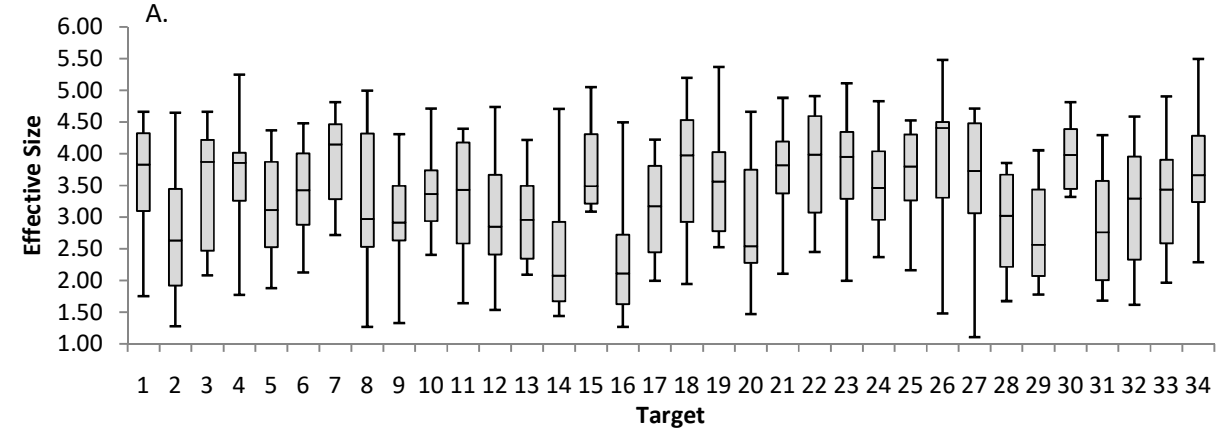
Example lineups used in the mock-witness task. Target-present lineups are presented on the left and corresponding target-absent lineups on the right.

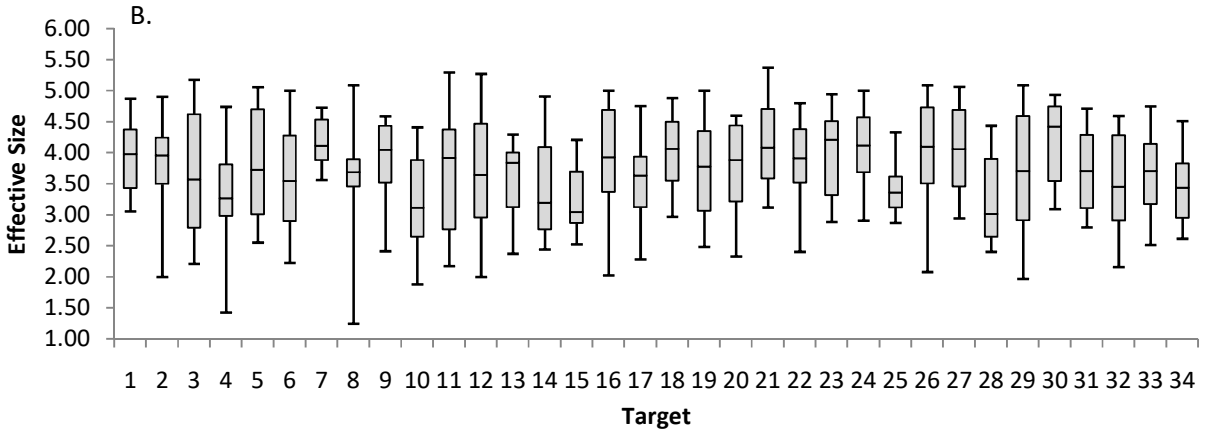


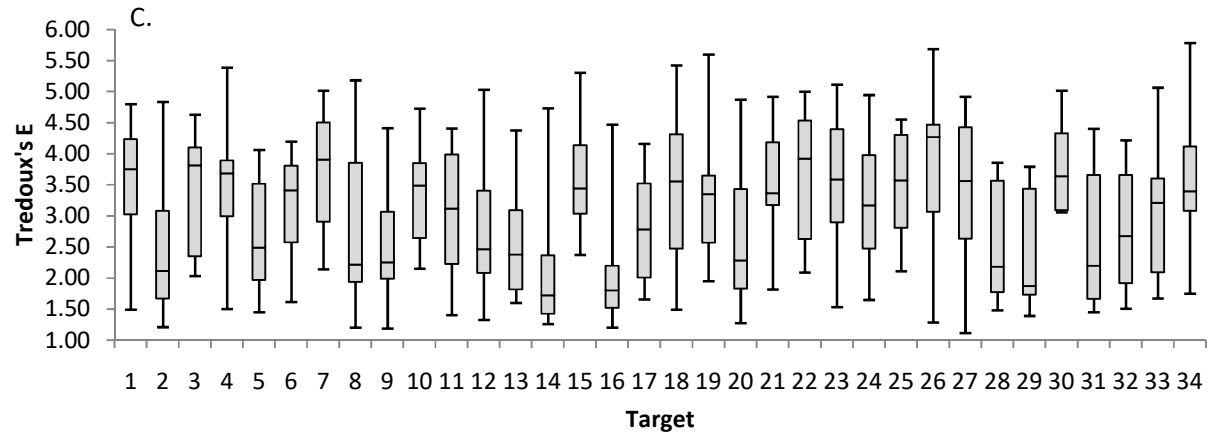
Note: The individuals whose faces appear here gave consent for the use of their likenesses.

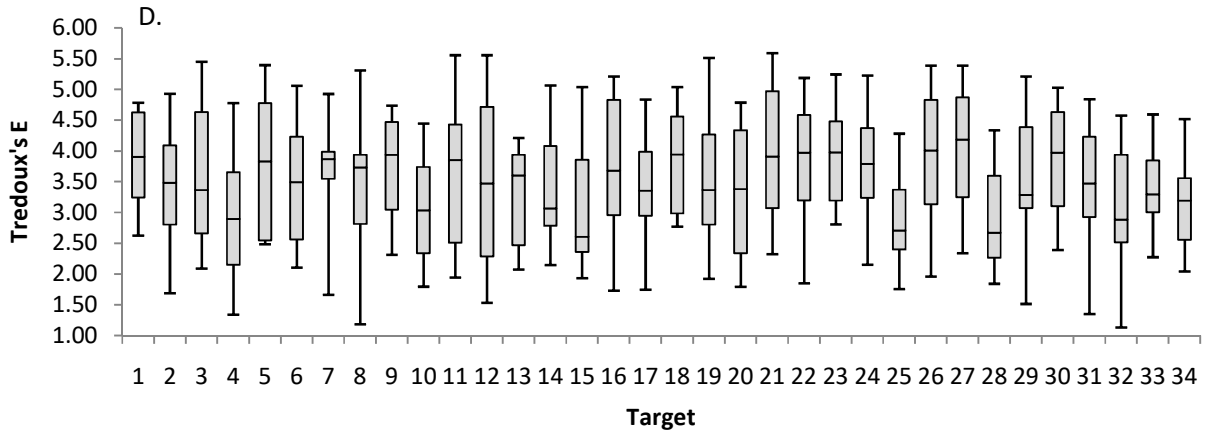
Supplemental Material 6

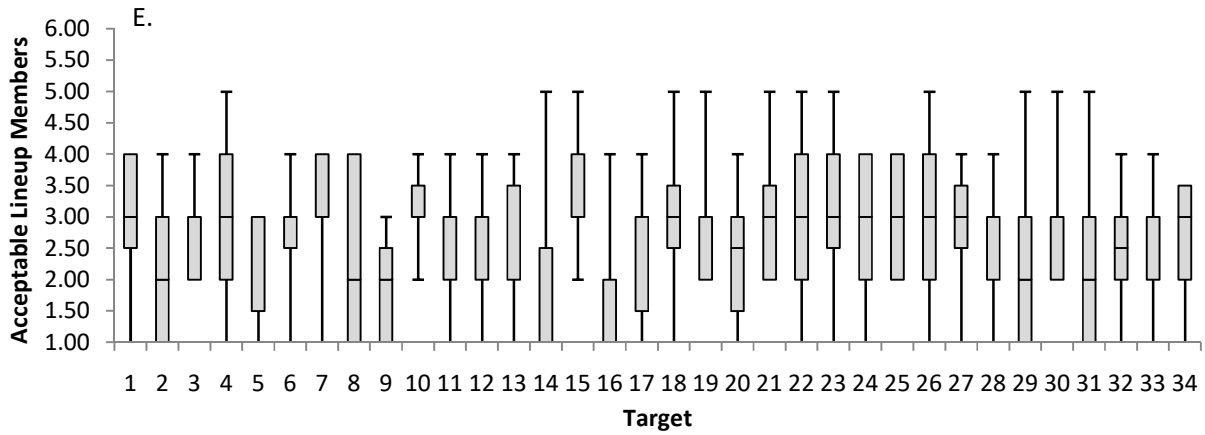
Boxplots representing the variability in lineup size measures by target. Panels A., C., and E. display results for target-present lineups and panels B., D., and F. display results for target-absent lineups (full data set).

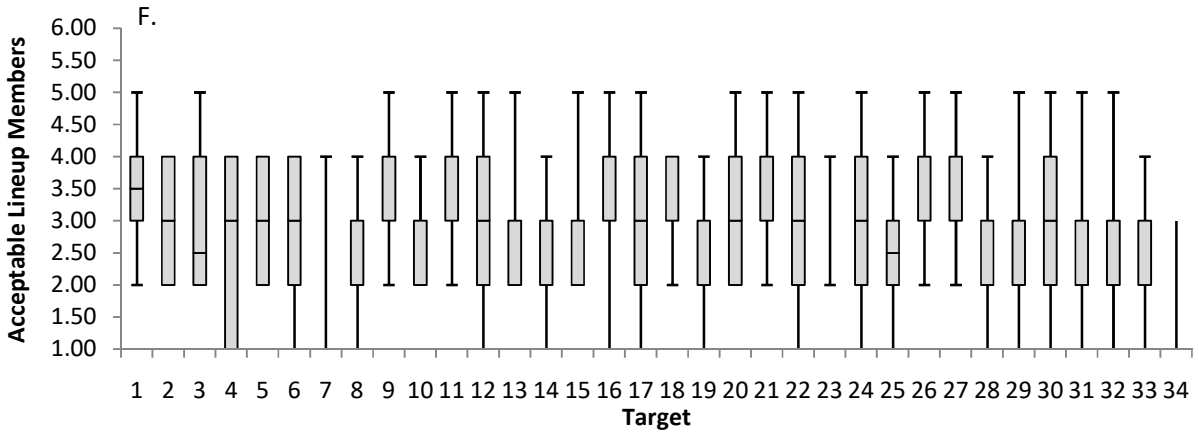






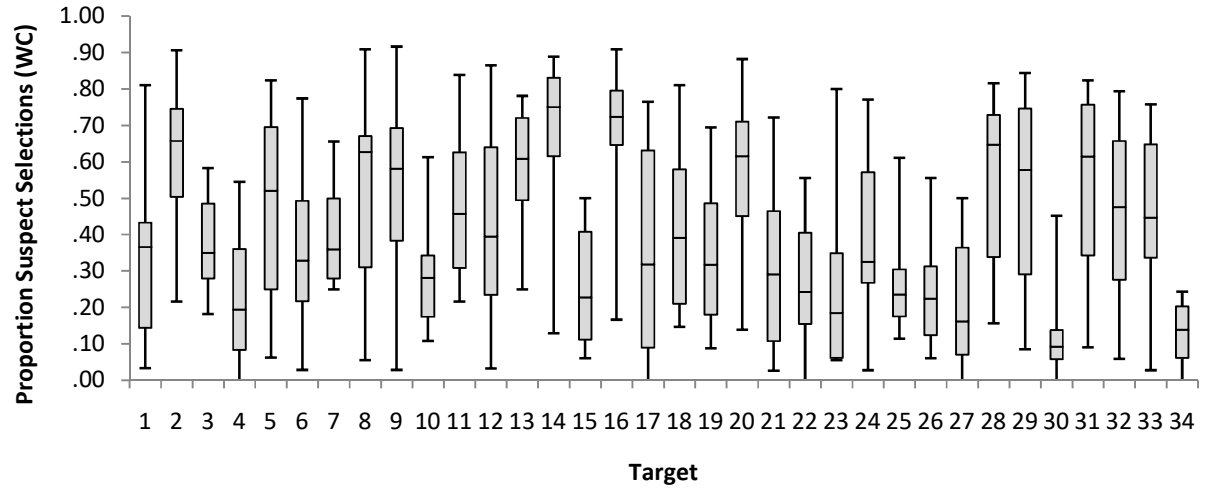

D.
Tredoux's E
6.00
5.50
5.00
4.50
4.00
3.50
3.00
2.50
2.00
1.50
1.00
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34
Target

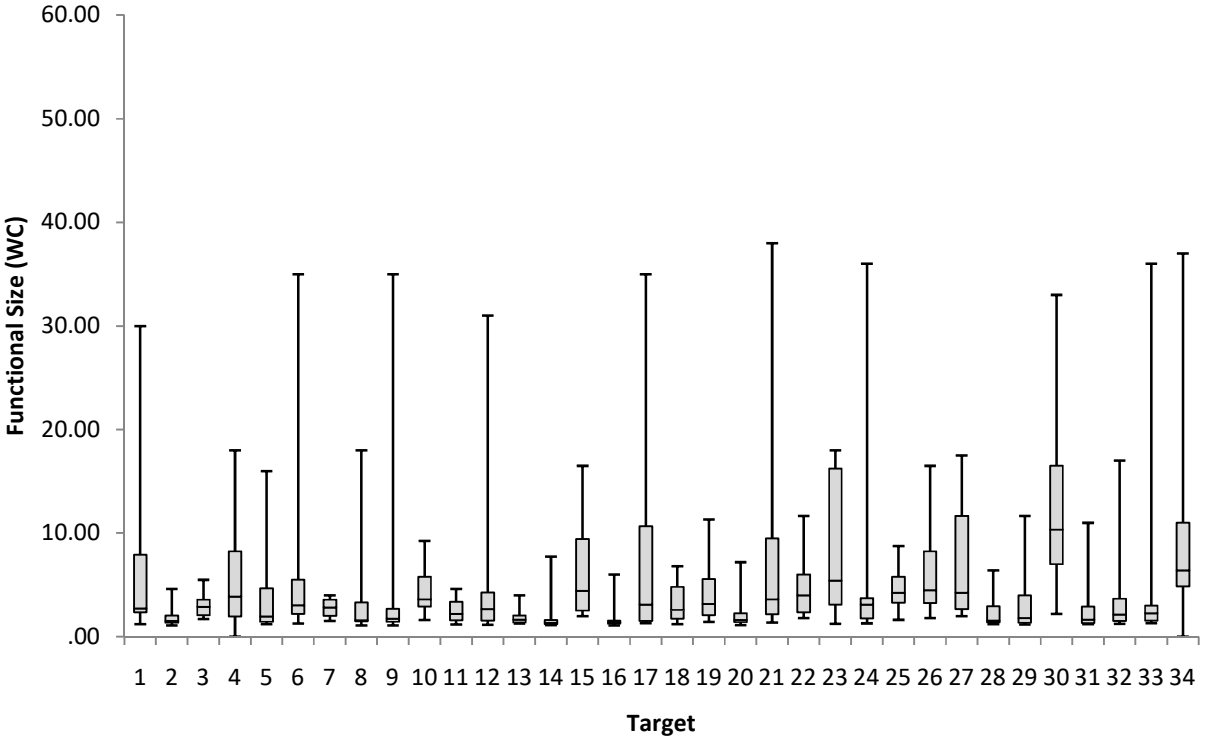

E.
Acceptable Lineup Members
6.00
5.50
5.00
4.50
4.00
3.50
3.00
2.50
2.00
1.50
1.00
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34
Target

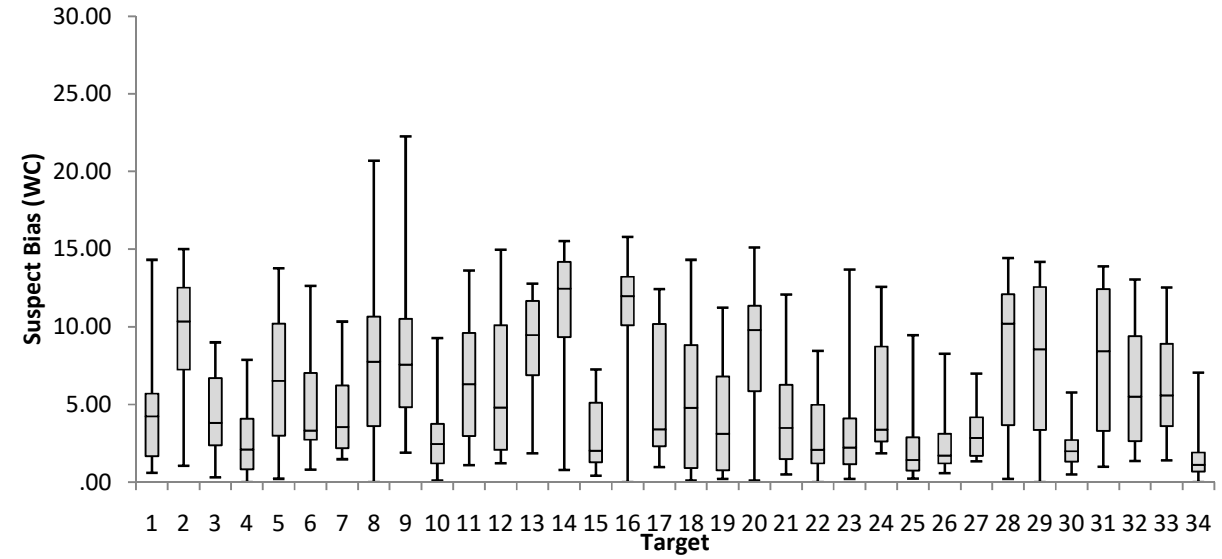

F.
Acceptable Lineup Members
6.00
5.50
5.00
4.50
4.00
3.50
3.00
2.50
2.00
1.50
1.00
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34
Target

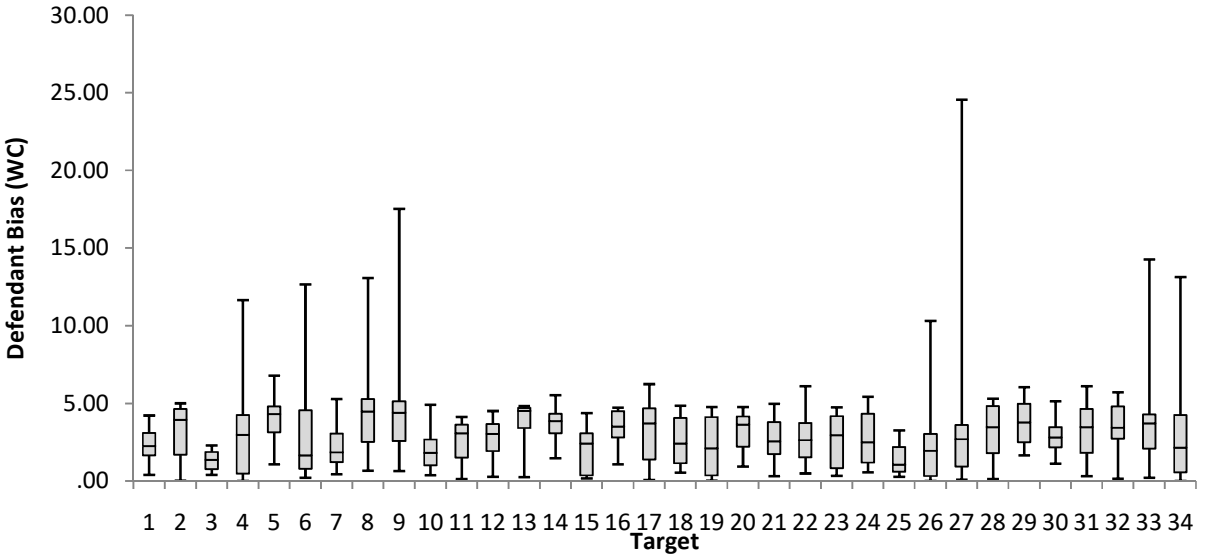
Supplemental Material 7

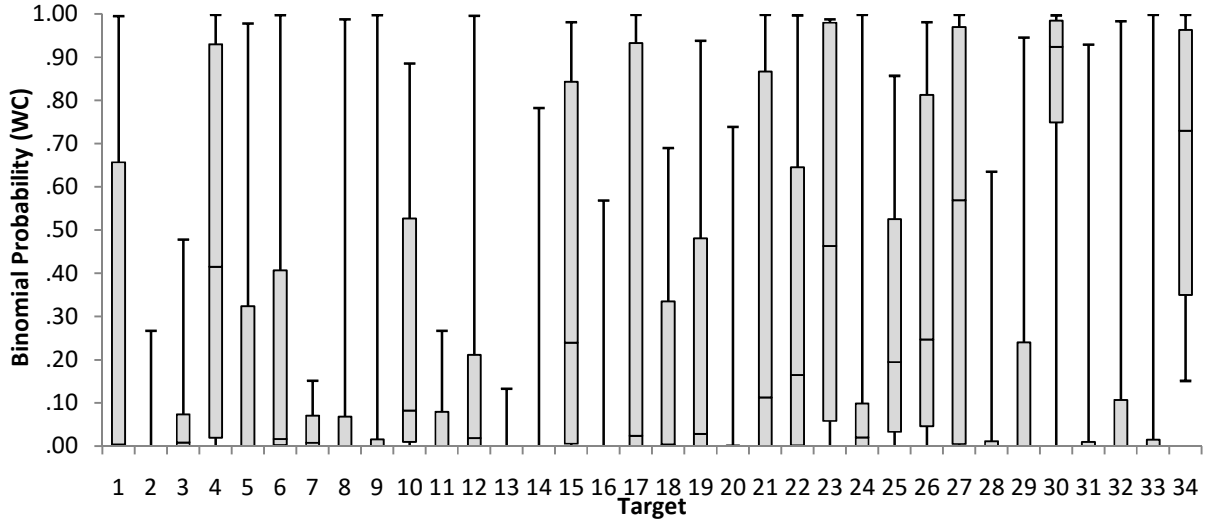
Boxplots representing the variability in lineup bias measures by target for target-present lineups (full data set) using a worst-case scenario approach to choose the innocent suspect.





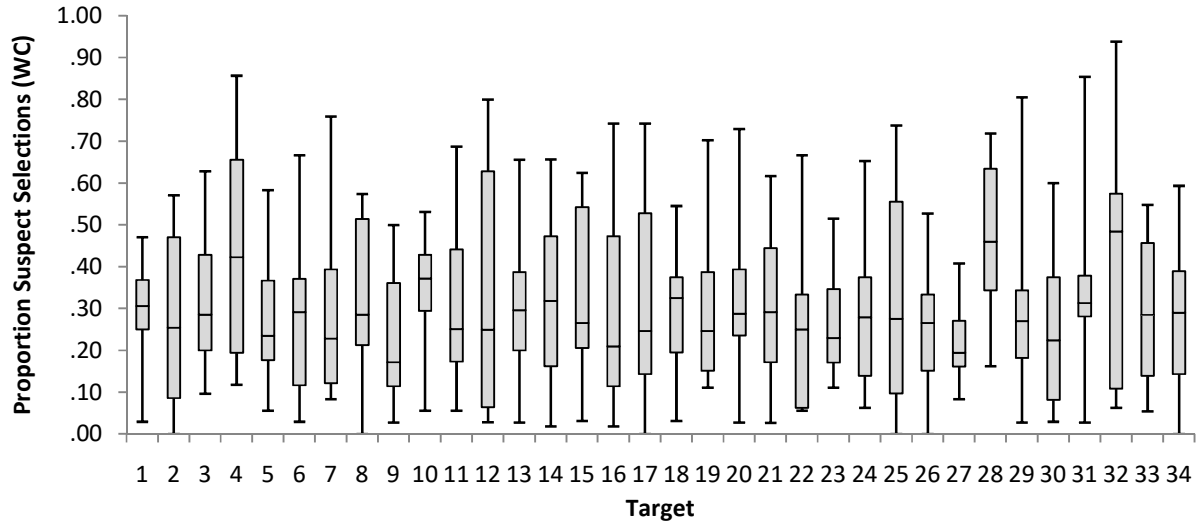

Suspect Bias (WC)
30.00
25.00
20.00
15.00
10.00
5.00
.00
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34
Target

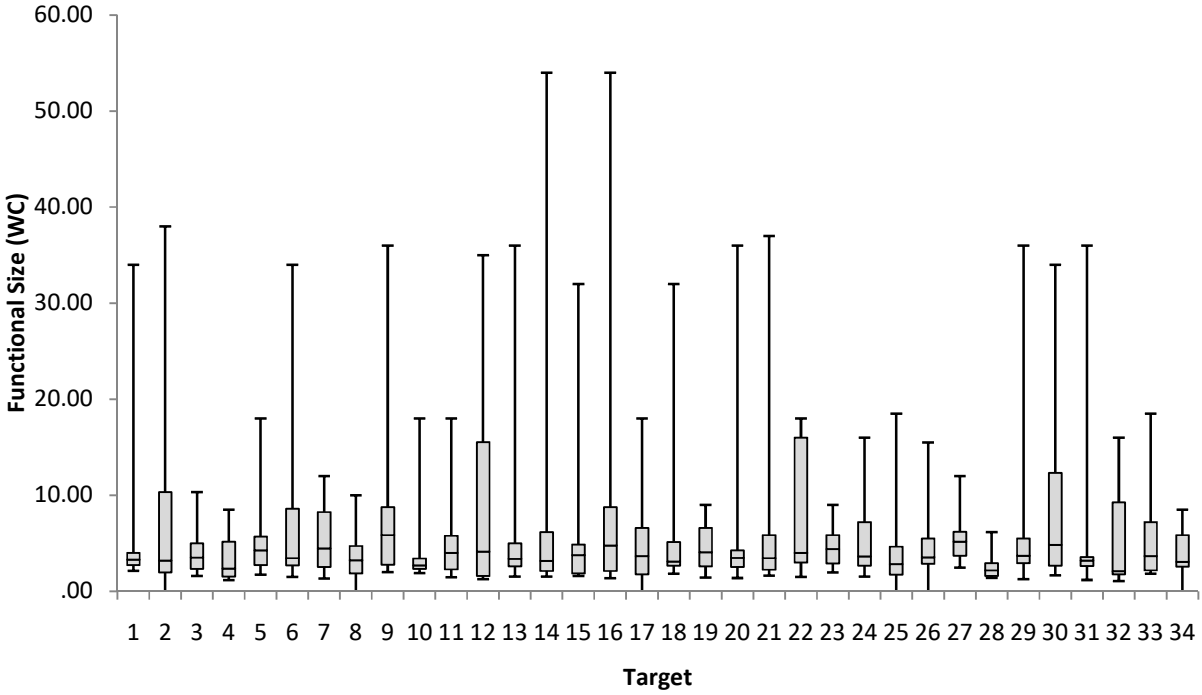

Defendant Bias (WC)
30.00
25.00
20.00
15.00
10.00
5.00
.00
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34
Target

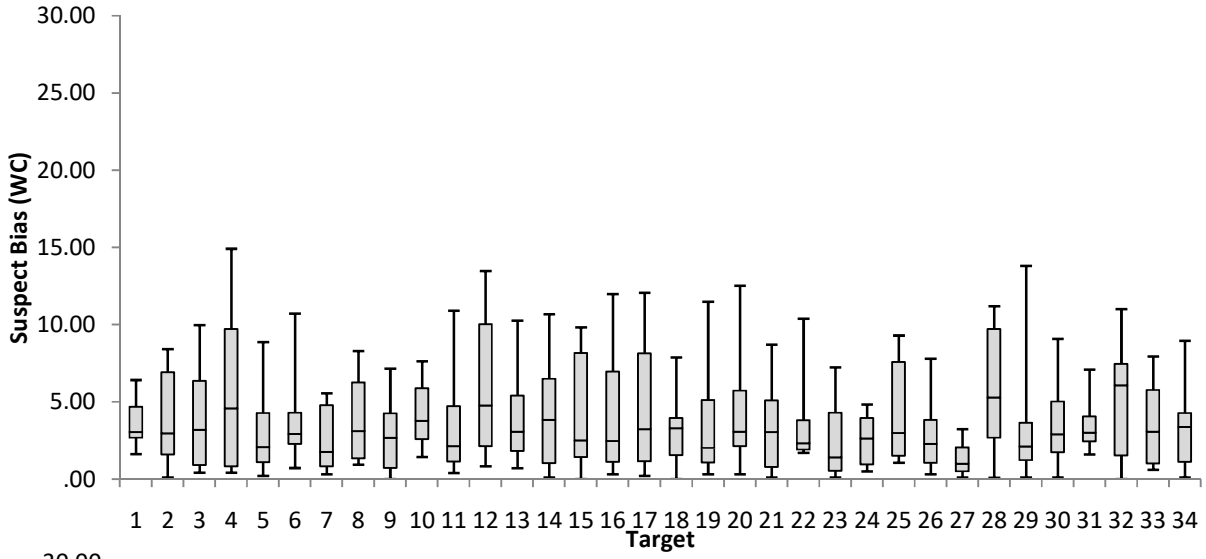

Binomial Probability (WC)
1.00
.90
.80
.70
.60
.50
.40
.30
.20
.10
.00
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34
Target

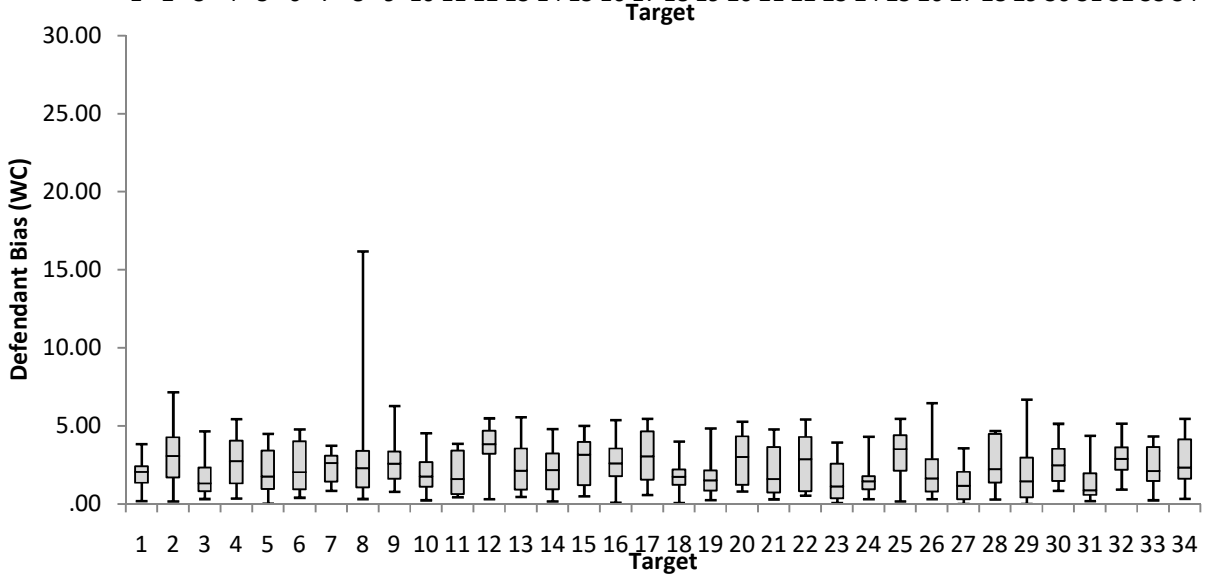
Supplemental Material 8

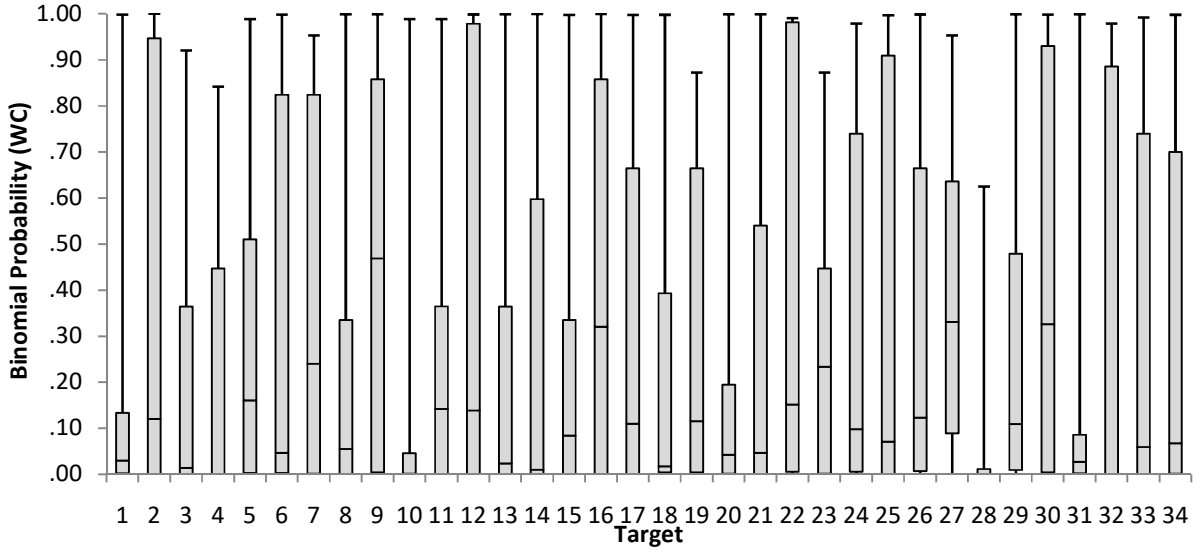
Boxplots representing the variability in lineup bias measures by target for target-absent lineups (suspect selected using the worst-case scenario approach).





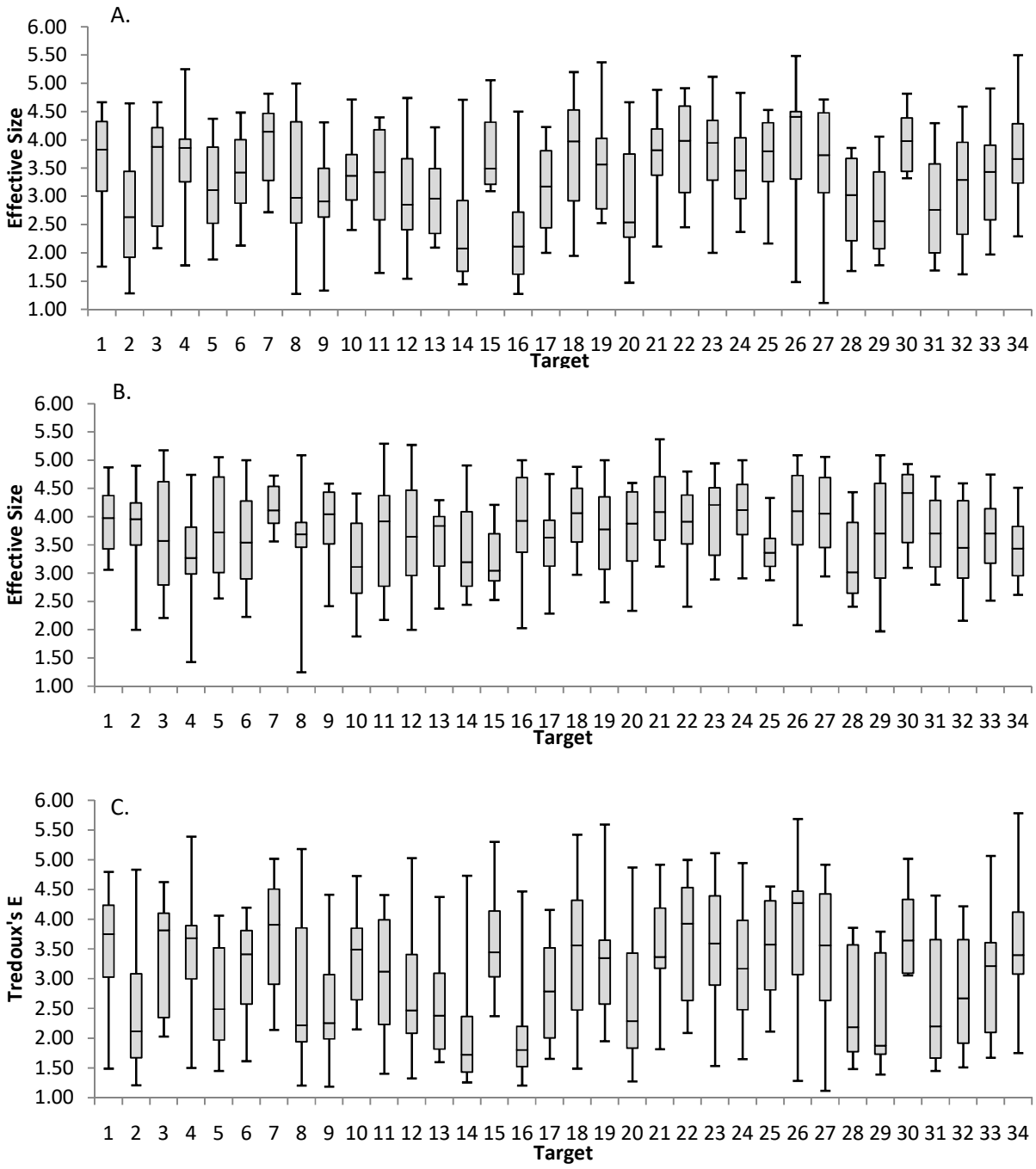

30.00
25.00
20.00
15.00
10.00
5.00
.00
Suspect Bias (WC)
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34
Target

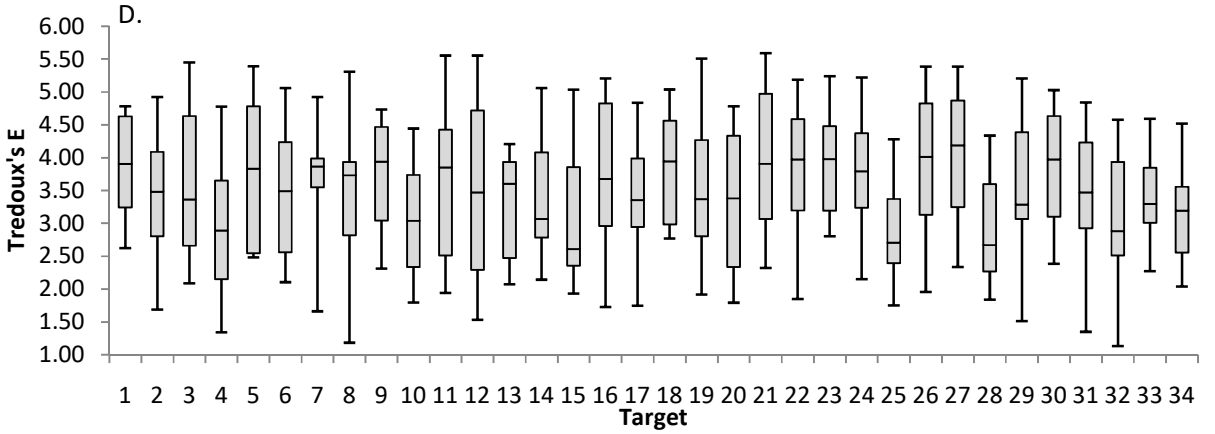

30.00
25.00
20.00
15.00
10.00
5.00
.00
Defendant Bias (WC)
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34
Target

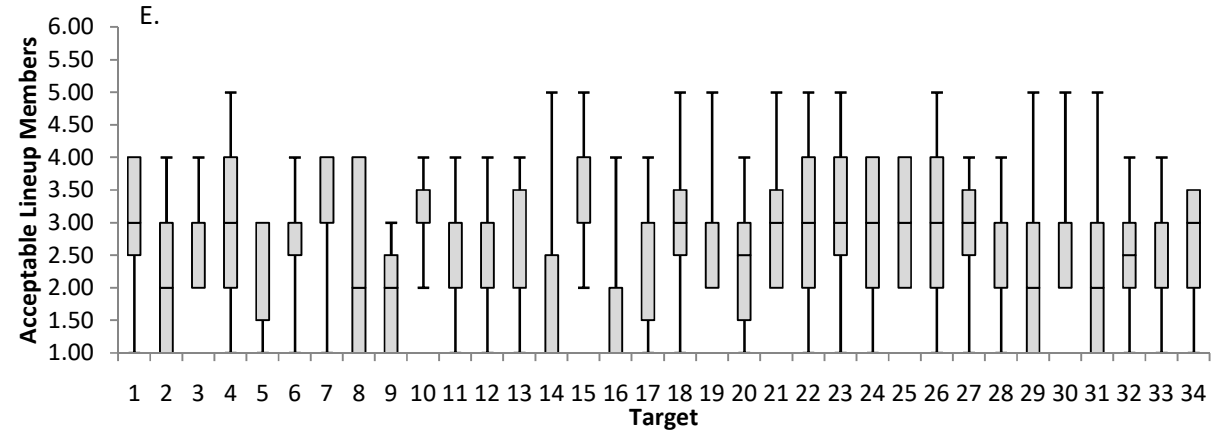

1.00
.90
.80
.70
.60
.50
.40
.30
.20
.10
.00
Binomial Probability (WC)
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34
Target

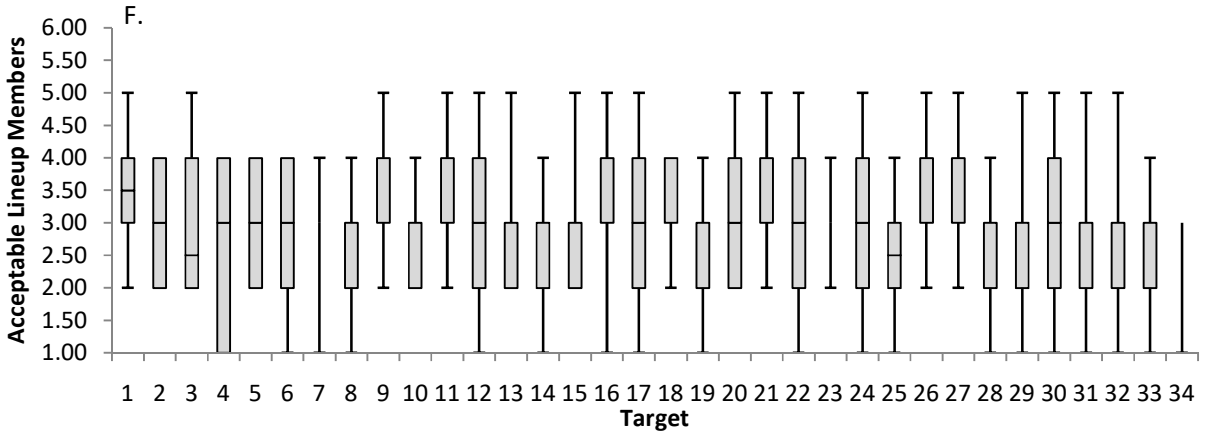
Supplemental Material 9

Boxplots representing the variability in lineup size measures by target. Panels A., C., and E. display results for target-present lineups and panels B., D., and F. display results for target-absent lineups (complete match data subset).



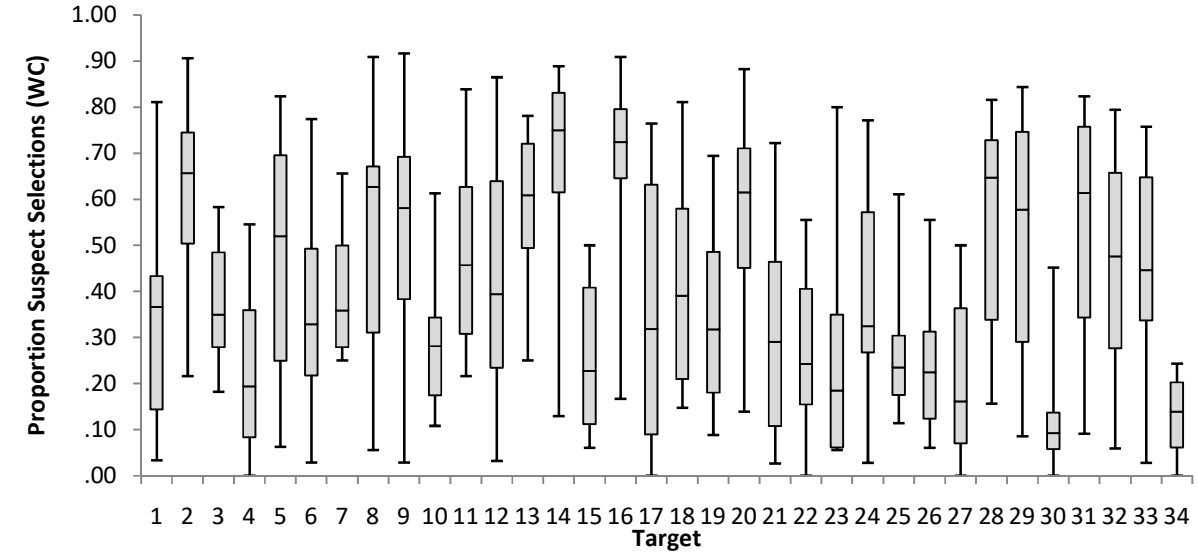

D.
Tredoux's E
6.00
5.50
5.00
4.50
4.00
3.50
3.00
2.50
2.00
1.50
1.00
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34
Target

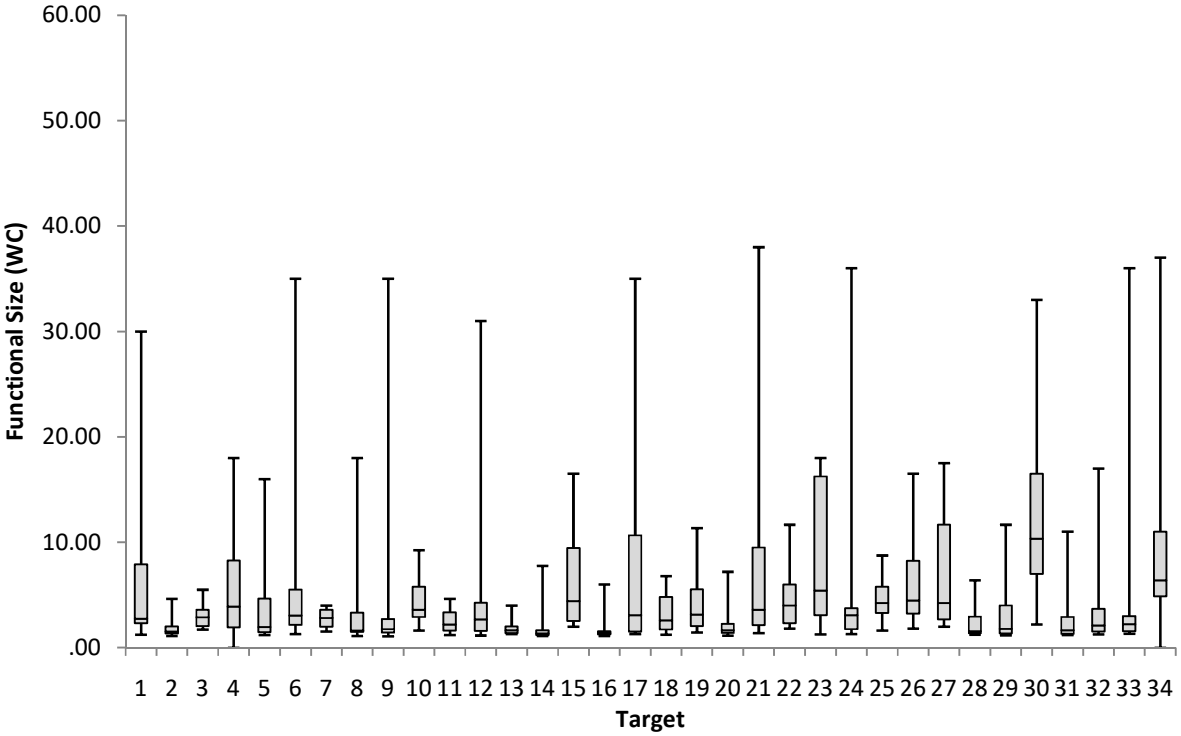

E.
Acceptable Lineup Members
6.00
5.50
5.00
4.50
4.00
3.50
3.00
2.50
2.00
1.50
1.00
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34
Target

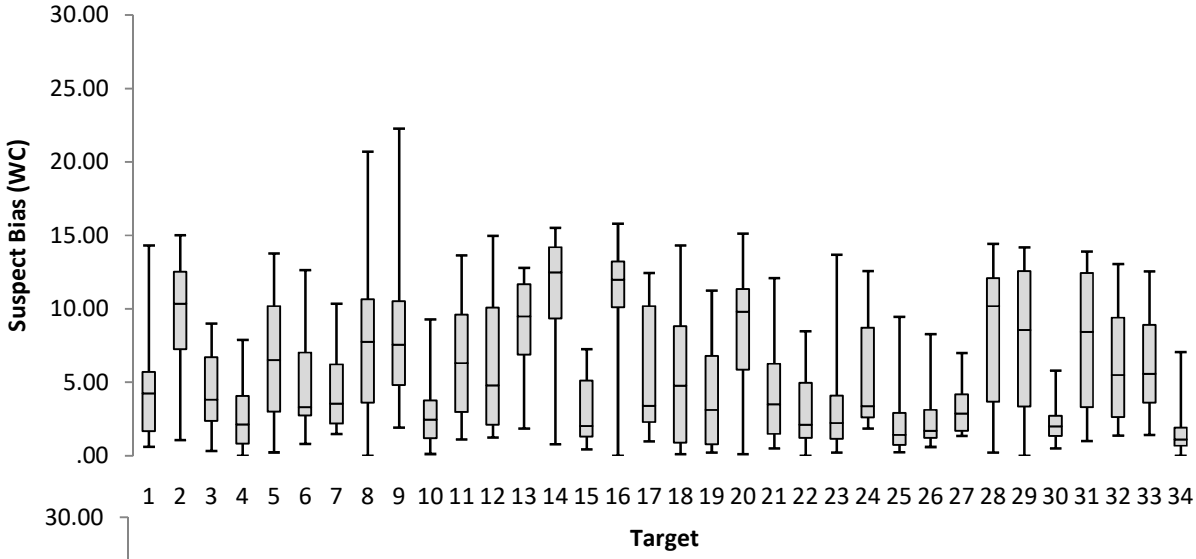

F.
Acceptable Lineup Members
6.00
5.50
5.00
4.50
4.00
3.50
3.00
2.50
2.00
1.50
1.00
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34
Target

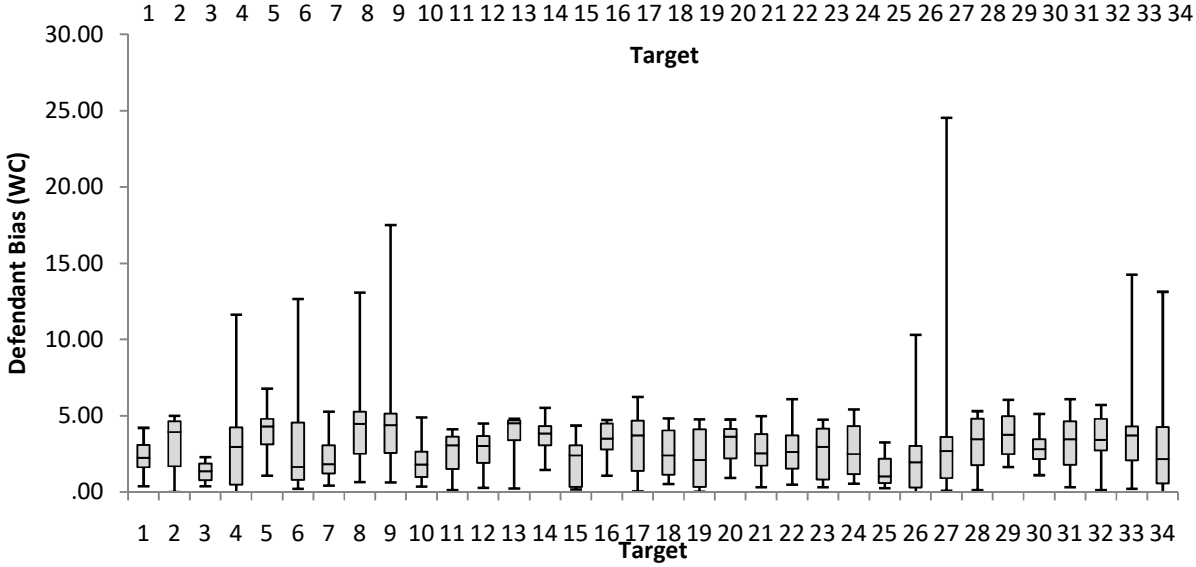
Supplemental Material 10

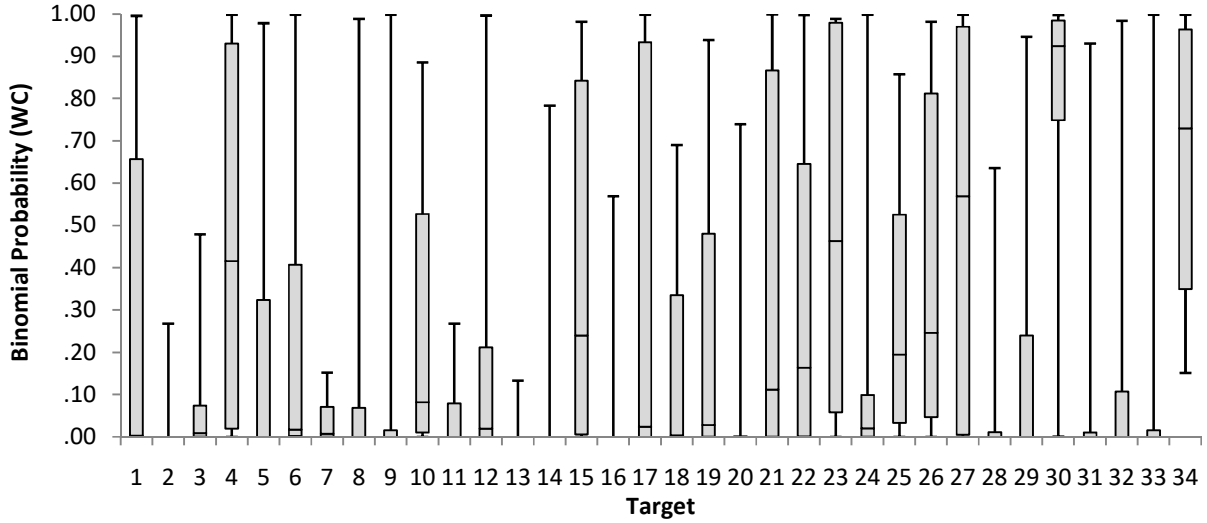
Boxplots representing the variability in lineup bias measures by target for target-present lineups (full data set).





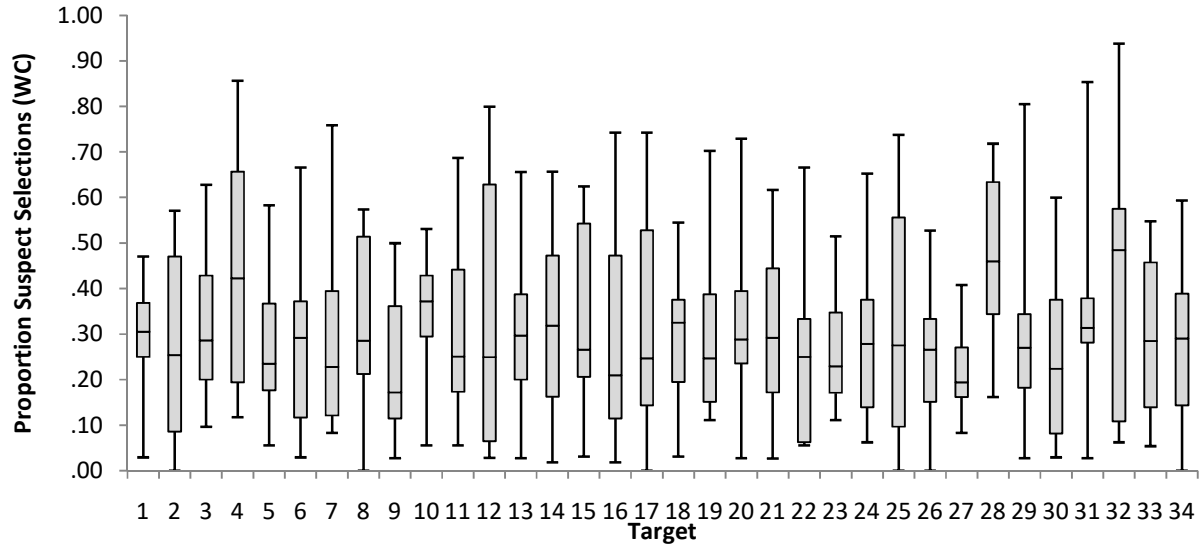

Suspect Bias (WC)
30.00
25.00
20.00
15.00
10.00
5.00
.00
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34
Target

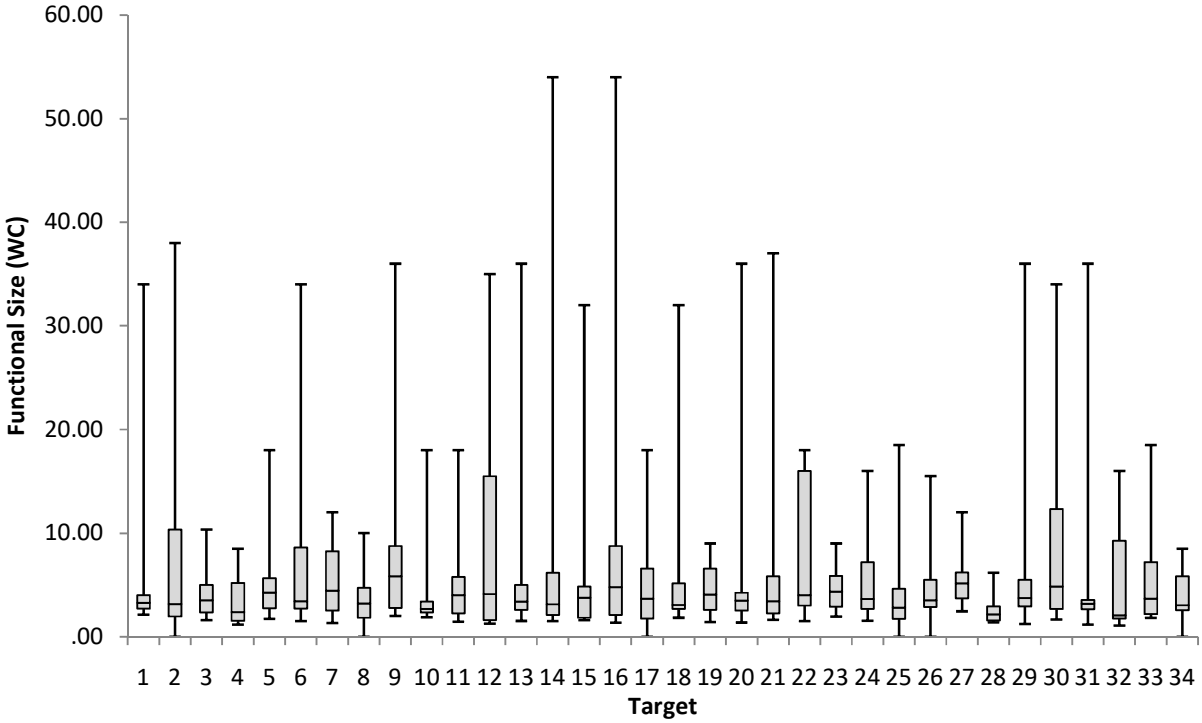

Defendant Bias (WC)
30.00
25.00
20.00
15.00
10.00
5.00
.00
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34
Target

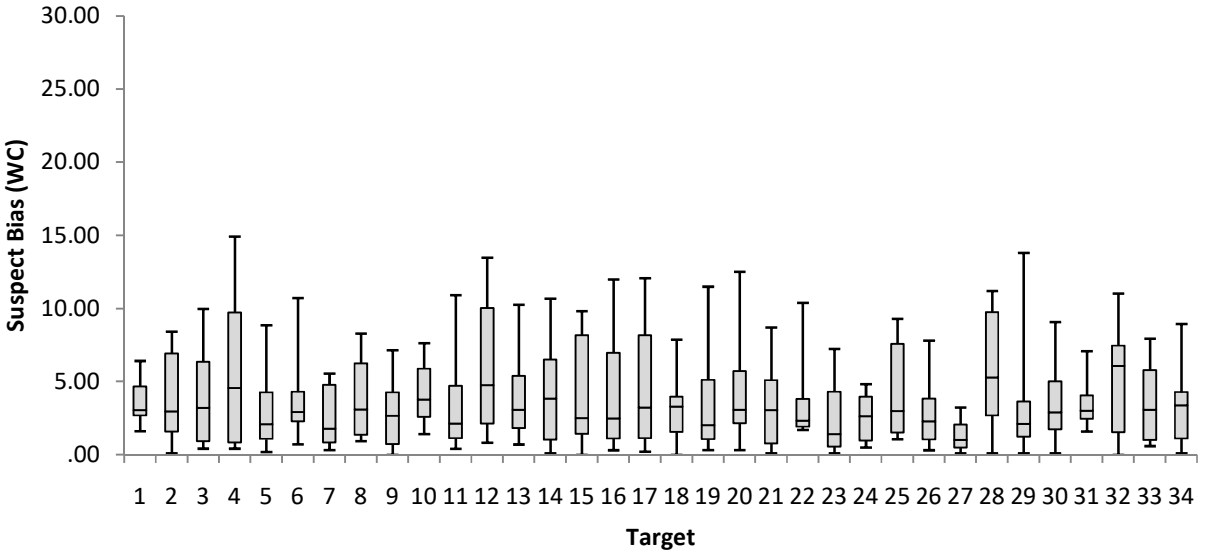

Binomial Probability (WC)
1.00
.90
.80
.70
.60
.50
.40
.30
.20
.10
.00
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34
Target

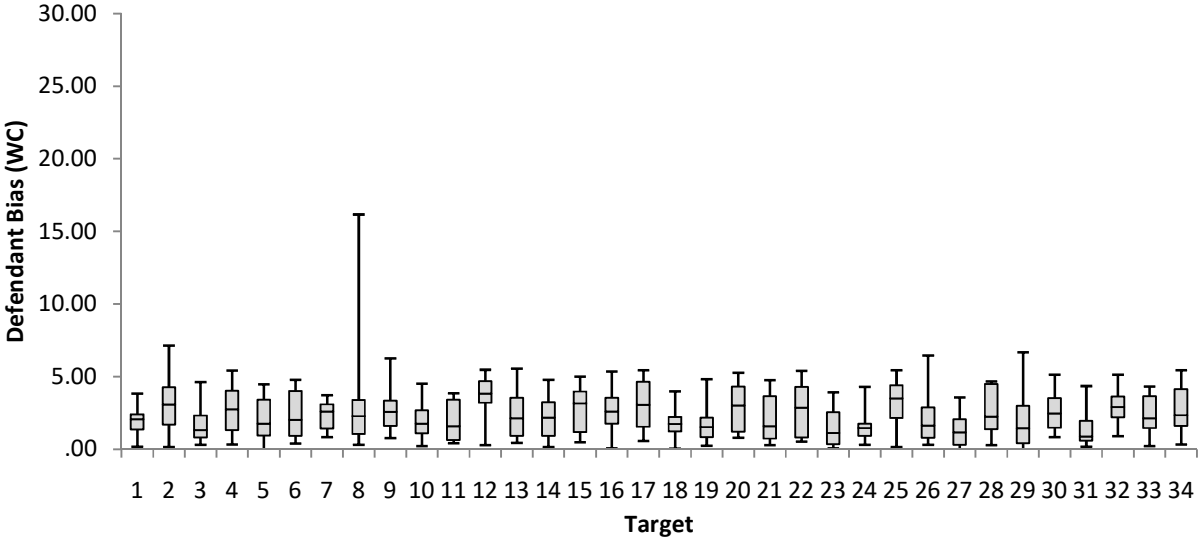
Supplemental Material 11

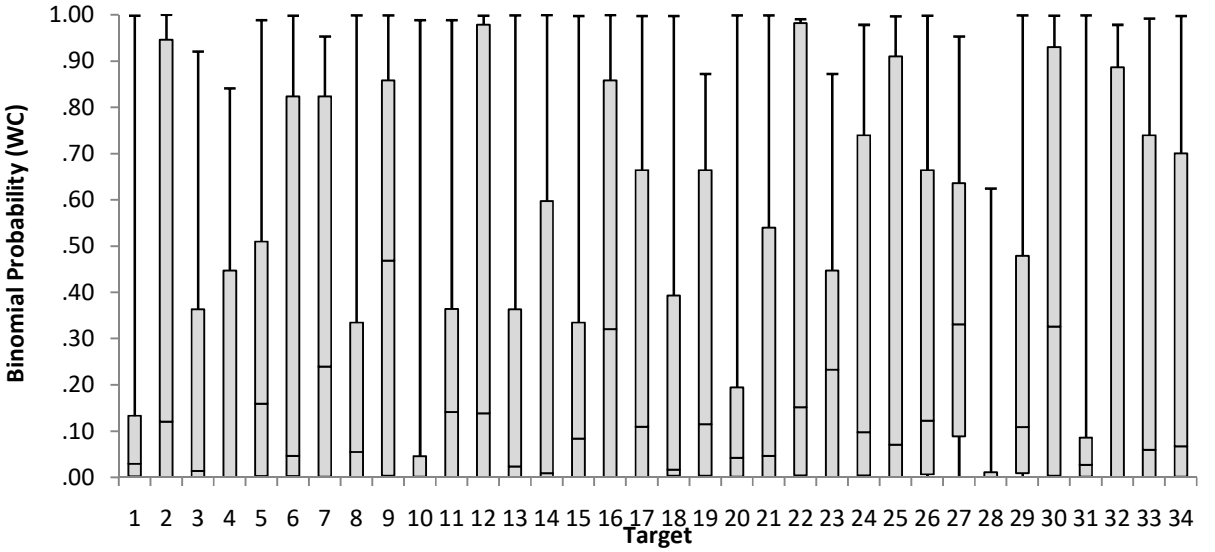
Boxplots representing the variability in lineup bias measures by target for target-absent lineups (suspect selected using the worst-case scenario approach).






30.00
25.00
20.00
15.00
10.00
5.00
.00
Suspect Bias (WC)
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34
Target


30.00
25.00
20.00
15.00
10.00
5.00
.00
Defendant Bias (WC)
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34
Target


1.00
.90
.80
.70
.60
.50
.40
.30
.20
.10
.00
Binomial Probability (WC)
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34
Target